Generalization Bounds for Domain Adaptation

Chao Zhang[1], Lei Zhang[2], Jieping Ye[1,3]
[1]Center for Evolutionary Medicine and Informatics, The Biodesign Institute,
and [3]Computer Science and Engineering, Arizona State University, Tempe, USA
zhangchao1015@gmail.com; jieping.ye@asu.edu
[2]School of Computer Science and Technology,
Nanjing University of Science and Technology, Nanjing, P.R. China
zhanglei.njust@yahoo.com.cn

April 8, 2013

Abstract

In this paper, we provide a new framework to obtain the generalization bounds of the learning process for domain adaptation, and then apply the derived bounds to analyze the asymptotical convergence of the learning process. Without loss of generality, we consider two kinds of representative domain adaptation: one is with multiple sources and the other is combining source and target data.

In particular, we use the integral probability metric to measure the difference between two domains. For either kind of domain adaptation, we develop a related Hoeffding-type deviation inequality and a symmetrization inequality to achieve the corresponding generalization bound based on the uniform entropy number. We also generalized the classical McDiarmid's inequality to a more general setting where independent random variables can take values from different domains. By using this inequality, we then obtain generalization bounds based on the Rademacher complexity. Afterwards, we analyze the asymptotic convergence and the rate of convergence of the learning process for such kind of domain adaptation. Meanwhile, we discuss the factors that affect the asymptotic behavior of the learning process and the numerical experiments support our theoretical findings as well.

1 Introduction

The generalization bound measures the probability that a function, chosen from a function class by an algorithm, has a sufficiently small error and plays an important role in statistical learning theory (see Vapnik, 1999; Bousquet *et al.*, 2004). The generalization bounds have been widely used to study the consistency of the ERM-based learning process (Vapnik, 1999), the asymptotic convergence of empirical process (Van der Vaart and Wellner, 1996) and the learnability of learning models (Blumer *et al.*, 1989). Generally, there are three essential aspects to obtain the generalization bounds of a specific learning process: complexity measures of function classes, deviation (or concentration) inequalities and symmetrization inequalities related to the learning process. For example, Van der Vaart and Wellner (1996) presented the generalization bounds

based on the Rademacher complexity and the covering number, respectively. Vapnik (1999) gave the generalization bounds based on the Vapnik-Chervonenkis (VC) dimension. Bartlett *et al.* (2005) proposed the local Rademacher complexity and obtained a sharp generalization bound for a particular function class $\{f \in \mathcal{F} : \mathrm{E}f^2 < \beta \mathrm{E}f, \beta > 0\}$. Hussain and Shawe-Taylor (2011) showed improved loss bounds for multiple kernel learning.

It is noteworthy that the aforementioned results of statistical learning theory are all built under the assumption that training and test data are drawn from the same distribution (or briefly called the assumption of same distribution). This assumption may not be valid in the situation that training and test data have different distributions, which will arise in many practical applications including speech recognition (Jiang and Zhai, 2007) and natural language processing (Blitzer *et al.*, 2007). Domain adaptation has recently been proposed to handle this situation and it is aimed to apply a learning model, trained by using the samples drawn from a certain domain (*source domain*), to the samples drawn from another domain (*target domain*) with a different distribution (see Bickel *et al.*, 2007; Wu and Dietterich, 2004; Blitzer *et al.*, 2006; Ben-David *et al.*, 2010; Bian *et al.*, 2012).

Without loss of generality, this paper is mainly concerned with two types of representative domain adaptation. In the first type, the learner receives training data from several source domains, known as *domain adaptation with multiple sources* (see Crammer *et al.*, 2006, 2008; Mansour *et al.*, 2008, 2009a). In the second type, the learner minimizes a convex combination of the source and the target empirical risks, termed as *domain adaptation combining source and target data* (see Ben-David *et al.*, 2010; Blitzer *et al.*, 2008).

1.1 Overview of Main Results

In this paper, we present a new framework to obtain the generalization bounds of the learning process for the aforementioned two kinds of representative domain adaptation, respectively. Based on the resultant bounds, we then analyze the asymptotical properties of the learning processes for the two types of domain adaptation. There are four major aspects in the framework:

- the quantity measuring the difference between two domains;
- the complexity measure of function class;
- the deviation inequalities of the learning process for domain adaptation;
- the symmetrization inequality of the learning precess for domain adaptation.

Generally, in order to obtain the generalization bounds of a learning process, one needs to develop the related deviation (or concentration) inequalities of the learning process. For either kind of domain adaptation, we use a martingale method to develop the related Hoeffding-type deviation inequality. Moreover, in the situation of domain adaptation, since the source domain differs from the target domain, the desired symmetrization inequality for domain adaptation should incorporate some quantity to reflect the difference. From the point of this view, we then obtain the related symmetrization inequality incorporating the integral probability metric that measures the difference between the distributions of the source and the target domains. Next, we present the generalization bounds based on the uniform entropy number for both kinds of domain adaptation. Also, we generalize the classical McDiarmid's inequality to a more general setting, where independent random variables take values from different domains. By using the derived

inequality, we obtain the generalization bounds based on the Rademacher complexity. Following the resultant bounds, we study the asymptotic convergence and the rate of convergence of the learning process in addition to a discussion on factors that affect the asymptotic behaviors. The numerical experiments support our theoretical findings as well. Meanwhile, we give a comparison with the related results under the assumption of same distribution.

1.2 Organization of the Paper

The rest of this paper is organized as follows. Section 2 introduces the problems studied in this paper. Section 3 introduces the integral probability metric to measure the difference between two domains. In Section 4, we introduce two kinds of complexity measures of function classes including the uniform entropy number and the Rademacher complexity. In Section 5 (resp. Section 6), we present the generalization bounds of the learning process for domain adaptation with multiple sources (resp. combining source and target data), and then analyze the asymptotic behavior of the learning process in addition to the related numerical experiment supporting our findings. In Section 7, we list the existing works on the theoretical analysis of domain adaptation as a comparison and the last section concludes the paper. In the appendices, we prove main results of this paper. For clarity of presentation, we also postpone the discussion of the deviation inequalities and the symmetrization inequalities in the appendices.

2 Problem Setup

In this section, we formalize the main issues of this paper by introducing some necessary notations

2.1 Domain Adaptation with Multiple Sources

We denote $\mathcal{Z}^{(S_k)} := \mathcal{X}^{(S_k)} \times \mathcal{Y}^{(S_k)} \subset \mathbb{R}^I \times \mathbb{R}^J$ $(1 \le k \le K)$ and $\mathcal{Z}^{(T)} := \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^I \times \mathbb{R}^J$ as the k-th source domain and the target domain, respectively. Set $L = I + J$. Let $\mathcal{D}^{(S_k)}$ and $\mathcal{D}^{(T)}$ stand for the distributions of the input spaces $\mathcal{X}^{(S_k)}$ $(1 \le k \le K)$ and $\mathcal{X}^{(T)}$, respectively. Denote $g_*^{(S_k)} : \mathcal{X}^{(S_k)} \to \mathcal{Y}^{(S_k)}$ and $g_*^{(T)} : \mathcal{X}^{(T)} \to \mathcal{Y}^{(T)}$ as the labeling functions of $\mathcal{Z}^{(S_k)}$ $(1 \le k \le K)$ and $\mathcal{Z}^{(T)}$, respectively. In the situation of domain adaptation with multiple sources, the input-space distributions $\mathcal{D}^{(S_k)}$ $(1 \le k \le K)$ and $\mathcal{D}^{(T)}$ differ from each other, or $g_*^{(S_k)}$ $(1 \le k \le K)$ and $g_*^{(T)}$ differ from each other, or both of the cases occur. There are sufficient amounts of i.i.d. samples $\mathbf{Z}_1^{N_k} = \{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}$ drawn from each source domain $\mathcal{Z}^{(S_k)}$ $(1 \le k \le K)$ but little or no labeled samples drawn from the target domain $\mathcal{Z}^{(T)}$.

Given $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, let $g_{\mathbf{w}} \in \mathcal{G}$ be the function that minimizes the empirical risk

$$\mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) = \sum_{k=1}^K w_k \mathrm{E}_{N_k}^{(S_k)}(\ell \circ g) = \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} \ell(g(\mathbf{x}_n^{(k)}), \mathbf{y}_n^{(k)}) \tag{1}$$

over $\mathcal{G}$ with respect to sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$, and it is expected that $g_{\mathbf{w}}$ will perform well on the target expected risk:

$$\mathrm{E}^{(T)}(\ell \circ g) := \int \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}), \quad g \in \mathcal{G}, \tag{2}$$

i.e., $g_{\mathbf{w}}$ approximates the labeling $g_*^{(T)}$ as precisely as possible.

In the learning process of domain adaptation with multiple sources, we are mainly interested in the following two types of quantities:

- $\mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}})$, which corresponds to the estimation of the expected risk in the target domain $\mathcal{Z}^{(T)}$ from the empirical quantity that is the weighted combination of the empirical risks in the multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$;
- $\mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*)$, which corresponds to the performance of the algorithm for domain adaptation with multiple sources,

where $\widetilde{g}_* \in \mathcal{G}$ is the function that minimizes the expected risk $\mathrm{E}^{(T)}(\ell \circ g)$ over $\mathcal{G}$.

Recalling (1) and (2), since

$$\mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ \widetilde{g}_*) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}}) \geq 0,$$

we have

$$\begin{aligned}
\mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) =& \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) + \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) \\
\leq& \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ \widetilde{g}_*) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}}) + \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{T}(\ell \circ \widetilde{g}_*) + \mathrm{E}^{T}(\ell \circ \widetilde{g}_*) \\
\leq& 2 \sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \right| + \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*),
\end{aligned} \tag{3}$$

and thus

$$0 \leq \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) \leq 2 \sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \right|.$$

This shows that the asymptotic behaviors of the aforementioned two quantities when the sample numbers $N_1, \cdots, N_K$ go to *infinity* can both be described by the supremum

$$\sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \right|, \tag{4}$$

which is the so-called generalization bound of the learning process for domain adaptation with multiple sources.

For convenience, we define the loss function as class

$$\mathcal{F} := \{\mathbf{z} \mapsto \ell(g(\mathbf{x}), \mathbf{y}) : g \in \mathcal{G}\}, \tag{5}$$

and call $\mathcal{F}$ as the function class in the rest of this paper. By (1) and (2), given sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ drawn from $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ respectively, we briefly denote for any $f \in \mathcal{F}$,

$$\mathrm{E}^{(T)} f := \int f(\mathbf{z}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}) \,, \tag{6}$$

and

$$\mathrm{E}_{\mathbf{w}}^{(S)} f := \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)}). \tag{7}$$

Thus, we rewrite the generalization bound (4) for domain adaptation with multiple sources as

$$\sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f \right|. \tag{8}$$

2.2 Domain Adaptation Combining Source and Target Data

Denote $\mathcal{Z}^{(S)} := \mathcal{X}^{(S)} \times \mathcal{Y}^{(S)} \subset \mathbb{R}^I \times \mathbb{R}^J$ and $\mathcal{Z}^{(T)} := \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^I \times \mathbb{R}^J$ as the source domain and the target domain, respectively. Let $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ stand for the distributions of the input spaces $\mathcal{X}^{(S)}$ and $\mathcal{X}^{(T)}$, respectively. Denote $g_*^{(S)} : \mathcal{X}^{(S)} \to \mathcal{Y}^{(S)}$ and $g_*^{(T)} : \mathcal{X}^{(T)} \to \mathcal{Y}^{(T)}$ as the labeling functions of $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. In the situation of domain adaptation combining source and target data (see Blitzer *et al.*, 2008; Ben-David *et al.*, 2010), the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ differ from each other, or the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$ differ from each other, or both cases occur. There are some (but not enough) samples $\mathbf{Z}_1^{N_T} := \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ drawn from the target domain $\mathcal{Z}^{(T)}$ in addition to a large amount of samples $\mathbf{Z}_1^{N_S} := \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ drawn from the source domain $\mathcal{Z}^{(S)}$ with $N^{(T)} \ll N^{(S)}$. Given a $\tau \in [0, 1)$, we denote $g_\tau \in \mathcal{G}$ as the function that minimizes the convex combination of the source and the target empirical risks over $\mathcal{G}$:

$$\mathrm{E}_\tau(\ell \circ g) := \tau \mathrm{E}_{N_T}^{(T)}(\ell \circ g) + (1 - \tau)\mathrm{E}_{N_S}^{(S)}(\ell \circ g), \tag{9}$$

and it is expected that g_τ will perform well for any pair $\mathbf{z}^{(T)} = (\mathbf{x}^{(T)}, \mathbf{y}^{(T)}) \in \mathcal{Z}^{(T)}$, *i.e.*, g_τ approximates the labeling function $g_*^{(T)}$ as precisely as possible.

As mentioned by Blitzer *et al.* (2008); Ben-David *et al.* (2010), setting τ involves a tradeoff between the source data that are sufficient but not accurate and the target data that are accurate but not sufficient. Especially, setting $\tau = 0$ provides a learning process of the basic domain adaptation with one single source (see Ben-David *et al.*, 2006).

Similar to the situation of domain adaptation with multiple sources, two types of quantities: $\mathrm{E}^{(T)}(\ell \circ g_\tau) - \mathrm{E}_\tau(\ell \circ g_\tau)$ and $\mathrm{E}^{(T)}(\ell \circ g_\tau) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*)$ also play an essential role in analyzing the asymptotic behavior of the learning process for domain adaptation combining source and target data. By the similar way of (3), we need to consider the supremum

$$\sup_{g \in \mathcal{G}} \left| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_\tau(\ell \circ g) \right|, \tag{10}$$

which is the so-called generalization bound of the learning process for domain adaptation combining source and target data. Following the notation of (5) and taking $f = \ell \circ g$, we can equivalently rewrite the generalization bound (10) as

$$\sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_\tau f \right|. \tag{11}$$

3 Integral Probability Metric

As shown in some existing works (see Mansour *et al.*, 2008, 2009a; Ben-David *et al.*, 2010, 2006), one of major challenges in the theoretical analysis of domain adaptation is to find a quantity to measure the difference between the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$. Then, one can use the quantity to achieve generalization bounds for domain adaptation. In this section, we use the integral probability metric to measure the difference between the distributions of $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, and then discuss the relationship between the integral probability metric and other quantities proposed in existing works, *e.g.*, the *$\mathcal{H}$-divergence* and the *discrepancy distance* (see Ben-David *et al.*, 2010; Mansour *et al.*, 2009b). Moreover, we will show that there is a special situation of domain adaptation, where the integral probability metric performs better than other quantities (see Remark 3.1)

3.1 Integral Probability Metric

In Ben-David *et al.* (2010, 2006), the $\mathcal{H}$-*divergence* was introduced to derive the generalization bounds based on the VC dimension under the condition of "λ-close". Mansour *et al.* (2009b) obtained the generalization bounds based on the Rademacher complexity by using the *discrepancy distance*. Both quantities are aimed to measure the difference between two input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$. Moreover, Mansour *et al.* (2009a) used the Rényi divergence to measure the distance between two distributions. In this paper, we use the following quantity to measure the difference between the distributions of the source and the target domains:

Definition 3.1 *Given two domains $\mathcal{Z}^{(S)}, \mathcal{Z}^{(T)} \subset \mathbb{R}^L$, let $\mathbf{z}^{(S)}$ and $\mathbf{z}^{(T)}$ be the random variables taking values from $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be a function class. We define*

$$D_{\mathcal{F}}(S,T) := \sup_{f\in\mathcal{F}} |\mathrm{E}^{(S)} f - \mathrm{E}^{(T)} f|, \tag{12}$$

where the expectations $\mathrm{E}^{(S)}$ and $\mathrm{E}^{(T)}$ are taken on the distributions $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively.

The quantity $D_{\mathcal{F}}(S,T)$ is termed as the integral probability metric that has played an important role in probability theory for measuring the difference between the two probability distributions (see Zolotarev, 1984; Rachev, 1991; Müller, 1997; Reid and Williamson, 2011). Recently, Sriperumbudur *et al.* (2009, 2012) gave the further investigation and proposed an empirical method to compute the integral probability metric. As mentioned by Müller (1997)[page 432], the quantity $D_{\mathcal{F}}(S,T)$ is a semimetric and it is a metric if and only if the function class $\mathcal{F}$ separates the set of all signed measures with $\mu(\mathcal{Z}) = 0$. Namely, according to Definition 3.1, given a non-trivial function class $\mathcal{F}$, the integral probability metric $D_{\mathcal{F}}(S,T)$ is equal to *zero* if the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$ have the same distribution.

By (5), the quantity $D_{\mathcal{F}}(S,T)$ can be equivalently rewritten as

$$\begin{aligned} D_{\mathcal{F}}(S,T) &= \sup_{g\in\mathcal{G}} \left| \mathrm{E}^{(S)} \ell(g(\mathbf{x}^{(S)}), \mathbf{y}^{(S)}) - \mathrm{E}^{(T)} \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) \right| \\ &= \sup_{g\in\mathcal{G}} \left| \mathrm{E}^{(S)} \ell\big(g(\mathbf{x}^{(S)}), g_*^{(S)}(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})\big) \right|. \end{aligned} \tag{13}$$

Next, based on the equivalent form (13), we discuss the relationships between the quantity $D_{\mathcal{F}}(S,T)$ and other quantities including the $\mathcal{H}$-*divergence* and the *discrepancy distance*.

3.2 Relationship with Other Quantities

Before the formal discussion, we briefly introduce the related quantities proposed in existing works (see Ben-David *et al.*, 2010; Mansour *et al.*, 2009b).

3.2.1 $\mathcal{H}$-Divergence and Discrepancy Distance

In classification tasks, by setting ℓ as the absolute-value loss function ($\ell(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|$), Ben-David *et al.* (2010) introduced a variant of the $\mathcal{H}$-*divergence*:

$$d_{\mathcal{H}\triangle\mathcal{H}}(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) = \sup_{g_1,g_2\in\mathcal{H}} \left| \mathrm{E}^{(S)} \ell\big(g_1(\mathbf{x}^{(S)}), g_2(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_2(\mathbf{x}^{(T)})\big) \right| \tag{14}$$

to achieve VC-dimension-based generalization bounds for domain adaptation under the condition of "λ-close": there exists a $\lambda > 0$ such that

$$\lambda \geq \inf_{g \in \mathcal{G}} \left\{ \int \ell(g(\mathbf{x}^{(S)}), \mathbf{y}^{(S)}) d\mathrm{P}(\mathbf{z}^{(S)}) + \int \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}) \right\}.$$

In both of the classification and regression tasks, given a function class $\mathcal{G}$ and a loss function ℓ, Mansour *et al.* (2009b) defined the *discrepancy distance* as

$$\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)}) = \sup_{g_1, g_2 \in \mathcal{G}} \left| \mathrm{E}^{(S)} \ell\big(g_1(\mathbf{x}^{(S)}), g_2(\mathbf{x}^{(S)})\big) - \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_2(\mathbf{x}^{(T)})\big) \right|, \tag{15}$$

and then used this quantity to obtain the generalization bounds based on the Rademacher complexity.

As mentioned by Mansour *et al.* (2009b), the quantities (14) and (15) match in the setting of classification tasks by setting ℓ as the absolute-value loss function, while the usage of (15) does not require the condition of "λ-close" but the usage of (14) does. Recalling Definition 3.1, since there is no limitation on the function class $\mathcal{F}$, the integral probability metric $D_{\mathcal{F}}(S, T)$ can be used in both classification and regression tasks. Therefore, we only consider the relationship between the integral probability metric $D_{\mathcal{F}}(S, T)$ and the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$.

3.2.2 Relationship between $D_{\mathcal{F}}(S, T)$ and $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$

From Definition 3.1 and (13), we can find that the integral probability metric $D_{\mathcal{F}}(S, T)$ measures the difference between the distributions of the two domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$. However, as addressed in Section 2, if a domain $\mathcal{Z}^{(S)}$ differs from another domain $\mathcal{Z}^{(T)}$, there are three possibilities: the input-space distribution $\mathcal{D}^{(S)}$ differs from $\mathcal{D}^{(T)}$, or $g_*^{(S)}$ differs from $g_*^{(T)}$, or both of them occur. Therefore, it is necessary to consider two kinds of differences: the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$. Next, we will show that the integral probability metric $D_{\mathcal{F}}(S, T)$ can be bounded by using two separate quantities that can measure the difference between $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

As shown in (15), the quantity $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ actually measures the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$. Moreover, we introduce another quantity to measure the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$:

Definition 3.2 *Given a loss function ℓ and a function class $\mathcal{G}$, we define*

$$Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)}) := \sup_{g_1 \in \mathcal{G}} \left| \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_*^{(T)}(\mathbf{x}^{(T)})\big) - \mathrm{E}^{(T)} \ell\big(g_1(\mathbf{x}^{(T)}), g_*^{(S)}(\mathbf{x}^{(T)})\big) \right|. \tag{16}$$

Note that if the loss function ℓ and the function class $\mathcal{G}$ are both non-trivial (*i.e.*, $\mathcal{F}$ is non-trivial), the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$ is a (semi)metric between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$. In fact, it is not hard to verify that $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$ satisfies the triangle inequality and is equal to *zero* if $g_*^{(S)}$ and $g_*^{(T)}$ match.

By combining (13), (15) and (16), we have

$$\begin{aligned}
\mathrm{disc}_\ell(\mathcal{D}^{(S)},\mathcal{D}^{(T)}) &= \sup_{g_1,g_2\in\mathcal{G}}\left|\mathrm{E}^{(S)}\ell\big(g_1(\mathbf{x}^{(S)}),g_2(\mathbf{x}^{(S)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_2(\mathbf{x}^{(T)})\big)\right| \\
&\geq \sup_{g_1\in\mathcal{G}}\left|\mathrm{E}^{(S)}\ell\big(g_1(\mathbf{x}^{(S)}),g_*^{(S)}(\mathbf{x}^{(S)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(S)}(\mathbf{x}^{(T)})\big)\right| \\
&= \sup_{g_1\in\mathcal{G}}\Big|\mathrm{E}^{(S)}\ell\big(g_1(\mathbf{x}^{(S)}),g_*^{(S)}(\mathbf{x}^{(S)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(T)}(\mathbf{x}^{(T)})\big) \\
&\quad +\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(T)}(\mathbf{x}^{(T)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(S)}(\mathbf{x}^{(T)})\big)\Big| \\
&\geq \sup_{g_1\in\mathcal{G}}\left|\mathrm{E}^{(S)}\ell\big(g_1(\mathbf{x}^{(S)}),g_*^{(S)}(\mathbf{x}^{(S)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(T)}(\mathbf{x}^{(T)})\big)\right| \\
&\quad -\sup_{g_1\in\mathcal{G}}\left|\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(T)}(\mathbf{x}^{(T)})\big)-\mathrm{E}^{(T)}\ell\big(g_1(\mathbf{x}^{(T)}),g_*^{(S)}(\mathbf{x}^{(T)})\big)\right| \\
&= D_{\mathcal{F}}(S,T)-Q_{\mathcal{G}}^{(T)}(g_*^{(S)},g_*^{(T)}),
\end{aligned} \tag{17}$$

and thus

$$D_{\mathcal{F}}(S,T)\leq \mathrm{disc}_\ell(\mathcal{D}^{(S)},\mathcal{D}^{(T)})+Q_{\mathcal{G}}^{(T)}(g_*^{(S)},g_*^{(T)}), \tag{18}$$

which implies that the integral probability metric $D_{\mathcal{F}}(S,T)$ can be bounded by the summation of the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)},\mathcal{D}^{(T)})$ and the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)},g_*^{(T)})$, which measure the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

Remark 3.1 *Note that there is a specific case in the situation of domain adaptation: $\mathcal{D}^{(S)}$ differs from $\mathcal{D}^{(T)}$ and meanwhile $g_*^{(S)}$ differs from $g_*^{(T)}$, while the distribution of the domain $\mathcal{Z}^{(S)}$ matches with that of the domain $\mathcal{Z}^{(T)}$. In this case, the integral probability metric $D_{\mathcal{F}}(S,T)$ equals to zero, but $\mathrm{disc}_\ell(\mathcal{D}^{(S)},\mathcal{D}^{(T)})$ or $Q_{\mathcal{G}}^{(T)}(g_*^{(S)},g_*^{(T)})$ neither equals to zero. Therefore, the integral probability metric $D_{\mathcal{F}}(S,T)$ is more suitable for this case than the discrepancy distance $\mathrm{disc}_\ell(\mathcal{D}^{(S)},\mathcal{D}^{(T)})$.*

4 Complexity Measures of Function Classes

Generally, the generalization bound of a certain learning process is achieved by incorporating some complexity measure of the function class, *e.g.*, the covering number, the VC dimension and the Rademacher complexity. In this paper, we are mainly concerned with the uniform entropy number and the Rademacher complexity.

4.1 Uniform Entropy Number

The uniform entropy number is derived from the concept of the covering number and we refer to Mendelson (2003) for details. The covering number of a function class $\mathcal{F}$ is defined as follows:

Definition 4.1 *Let $\mathcal{F}$ be a function class and d be a metric on $\mathcal{F}$. For any $\xi>0$, the covering number of $\mathcal{F}$ at radius ξ with respect to the metric d, denoted by $\mathcal{N}(\mathcal{F},\xi,d)$ is the minimum size of a cover of radius ξ.*

In some classical results of statistical learning theory, the covering number is applied by letting d be the distribution-dependent metric. For example, as shown in Theorem 2.3 of Mendelson (2003), one can set d as the norm $\ell_1(\mathbf{Z}_1^N)$ and then derive the generalization bound of the i.i.d. learning process by incorporating the expectation of the covering number, *i.e.,* $\mathrm{E}\mathcal{N}(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N))$. However, in the situation of domain adaptation, we only know the information of the source domain, while the expectation $\mathrm{E}\mathcal{N}(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N))$ is dependent on the distributions of the source and the target domains because $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Therefore, the covering number is no longer suitable for our framework to obtain the generalization bounds for domain adaptation. In contrast, the uniform entropy number is distribution-free and thus we choose it as the complexity measure of function classes to derive the generalization bounds for domain adaptation.

Next, we will consider the uniform entropy number of $\mathcal{F}$ in the situations of two types of domain adaptation: (i) domain adaptation with multiple sources; (ii) domain adaptation combining source and target data, respectively.

4.1.1 Domain Adaptation with Multiple Sources

For clarity of presentation, we give a useful notation for the following discussion. Let $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K := \{\{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}\}_{k=1}^K$ be the collection of sample sets drawn from multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$, respectively. Denote $\{\mathbf{Z'}_1^{N_k}\}_{k=1}^K := \{\{\mathbf{z'}_n^{(k)}\}_{n=1}^{N_k}\}_{k=1}^K$ as the collection of the ghost sample sets drawn from $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ such that the ghost sample $\mathbf{z'}_n^{(k)}$ has the same distribution as $\mathbf{z}_n^{(k)}$ for any $1 \leq k \leq K$ and any $1 \leq n \leq N_k$. Denote $\mathbf{Z}_1^{2N_k} := \{\mathbf{Z}_1^{N_k}, \mathbf{Z'}_1^{N_k}\}$ for any $1 \leq k \leq K$. Moreover, given an $f \in \mathcal{F}$ and a $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, we introduce a variant of the ℓ_1 norm:

$$\|f\|_{\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)} := \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} |f(\mathbf{z}_n^{(k)})|. \tag{19}$$

It is noteworthy that the variant $\ell_1^{\mathbf{w}}$ of the ℓ_1 norm is still a norm on the functional space, which can be directly verified by using the definition of norm, so we omit it here.

In the situation of domain adaptation with multiple sources, by setting the metric d as $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$, we then define the uniform entropy number of $\mathcal{F}$ with respect to the metric $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$ as

$$\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi, 2\sum_{k=1}^K N_k\big) := \sup_{\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K} \ln \mathcal{N}\left(\mathcal{F}, \xi, \ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right). \tag{20}$$

4.1.2 Domain Adaptation Combining Source and Target Data

In the situation of domain adaptation combining source and target data, we have to introduce another variant of the ℓ_1 norm on $\mathcal{F}$. Let $\mathbf{Z}_1^{N_S} = \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} = \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be two sets of samples drawn from the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Given an $f \in \mathcal{F}$, we define for any $\tau \in [0,1)$,

$$\|f\|_{\ell_1^\tau(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T})} := \frac{\tau}{N_T} \sum_{n=1}^{N_T} |f(\mathbf{z}_n^{(T)})| + \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} |f(\mathbf{z}_n^{(S)})|. \tag{21}$$

Note that the variant ℓ_1^τ ($\tau \in [0,1)$) of the norm ℓ_1 is still a norm on the functional space, which can be easily verified by using the definition of norm, so we omit it here.

Moreover, let $\mathbf{Z}'^{N_S}_1$ and $\overline{\mathbf{Z}}'^{N_T}_1$ be the ghost sample sets of $\mathbf{Z}^{N_S}_1$ and $\overline{\mathbf{Z}}^{N_T}_1$, respectively. Denote $\mathbf{Z}^{2N_S}_1 := \{\mathbf{Z}^{N_S}_1, \mathbf{Z}'^{N_S}_1\}$ and $\overline{\mathbf{Z}}^{2N_T}_1 := \{\overline{\mathbf{Z}}^{N_T}_1, \overline{\mathbf{Z}}'^{N_T}_1\}$, respectively. Then, the uniform entropy number of $\mathcal{F}$ with respect to the metric $\ell^\tau_1(\mathbf{Z})$ is defined as

$$\ln \mathcal{N}^\tau_1(\mathcal{F}, \xi, 2(N_S + N_T)) := \sup_{\mathbf{Z}} \ln \mathcal{N}\left(\mathcal{F}, \xi, \ell^\tau_1(\mathbf{Z})\right), \tag{22}$$

where $\mathbf{Z} := \{\mathbf{Z}^{2N_S}_1, \overline{\mathbf{Z}}^{2N_T}_1\}$.

4.2 Rademacher Complexity

The Rademacher complexity is one of the most frequently used complexity measures of function classes and we refer to Van der Vaart and Wellner (1996); Mendelson (2003) for details.

Definition 4.2 *Let $\mathcal{F}$ be a function class and $\{\mathbf{z}_n\}^N_{n=1}$ be a sample set drawn from $\mathcal{Z}$. Denote $\{\sigma_n\}^N_{n=1}$ be a set of random variables independently taking either value from $\{-1, 1\}$ with equal probability. Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}(\mathcal{F}) := \mathrm{E} \sup_{f\in\mathcal{F}} \left\{ \frac{1}{N} \Big| \sum_{n=1}^N \sigma_n f(\mathbf{z}_n) \Big| \right\} \tag{23}$$

with its empirical version

$$\mathcal{R}_N(\mathcal{F}) := \mathrm{E}_\sigma \sup_{f\in\mathcal{F}} \left\{ \frac{1}{N} \Big| \sum_{n=1}^N \sigma_n f(\mathbf{z}_n) \Big| \right\}, \tag{24}$$

where E *stands for the expectation taken with respect to all random variables $\{\mathbf{z}_n\}^N_{n=1}$ and $\{\sigma_n\}^N_{n=1}$, and* E_σ *stands for the expectation only taken with respect to the random variables $\{\sigma_n\}^N_{n=1}$.*

5 Learning Processes of Domain Adaptation with Multiple Sources

In this section, we present two generalization bounds of the learning process for domain adaptation with multiple sources. They are based on the uniform entropy number and the Rademacher complexity, respectively. By using the derived bounds based on the uniform entropy number, we then analyze the asymptotic convergence and the rate of convergence of the learning process. The numerical experiment supports our theoretical analysis as well.

5.1 Generalization Bounds

Based on the uniform entropy number defined in (20), a generalization bound for domain adaptation with multiple sources is presented in the following theorem.

Theorem 5.1 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a,b]$. Let $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$. Then, given an arbitrary $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, we have for any $\left(\prod_{k=1}^K N_k\right) \geq \frac{8(b-a)^2}{(\xi')^2}$ and any $\epsilon > 0$, with probability at least $1-\epsilon$,*

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_{\mathbf{w}}^{(S)} f - \mathrm{E}^{(T)} f\right| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left(\frac{\left(\ln \mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k\right) - \ln(\epsilon/8)\right)}{\frac{\left(\prod_{k=1}^K N_k\right)}{32(b-a)^2\left(\sum_{k=1}^K w_k^2 (\prod_{i\neq k} N_i)\right)}} \right)^{\frac{1}{2}}, \tag{25}$$

where $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and

$$D_{\mathcal{F}}^{(\mathbf{w})}(S,T) := \sum_{k=1}^K w_k D_{\mathcal{F}}(S_k, T). \tag{26}$$

In the above theorem, we show that the generalization bound $\sup_{f\in\mathcal{F}} |\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f|$ can be bounded by the right-hand side of (25). Compared to the classical result under the assumption of same distribution (see Mendelson, 2003, Theorem 2.3 and Definition 2.5): with probability at least $1-\epsilon$,

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_N f - \mathrm{E} f\right| \leq O\left(\left(\frac{\ln \mathcal{N}_1\left(\mathcal{F}, \xi, N\right) - \ln(\epsilon/8)}{N} \right)^{\frac{1}{2}} \right) \tag{27}$$

with $\mathrm{E}_N f$ being the empirical risk with respect to the sample set $\mathbf{Z}_1^N$, there is a discrepancy quantity $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ that is determined by the two factors: the choice of $\mathbf{w}$ and the integral probability metrics $D_{\mathcal{F}}(S_k,T)$ $(1 \leq k \leq K)$. The two results will coincide if any source domain and the target domain match, *i.e.*, $D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$.

In order to prove this result, we develop the specific Hoeffding-type deviation inequality and the symmetrization inequality for domain adaptation with multiple sources, respectively. The detailed proof is arranged in Appendix A. Subsequently, we give another generalization bound based on the Rademacher complexity:

Theorem 5.2 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a,b]$. Let $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$. Then, we have with probability at least $1-\epsilon$,*

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_{\mathbf{w}}^{(S)} f - \mathrm{E}^{(T)} f\right| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + 2\sum_{k=1}^K w_k \mathcal{R}^{(k)}(\mathcal{F}) + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}}, \tag{28}$$

where $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ is defined in (26) and $\mathcal{R}^{(k)}(\mathcal{F})$ $(1 \leq k \leq K)$ are the Rademacher complexities on the source domains $\mathcal{Z}^{(S_k)}$, respectively.

Similarly, the derived bound (28) coincides with the related classical result under the assumption of same distribution (see Bousquet *et al.*, 2004, Theorem 5), when any source domain of

$\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.*, $D_{\mathcal{F}}^{(\mathbf{w})}(S,T) = D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$. The proof of this theorem is processed by introducing a generalized version of McDiarmid's inequality which allows independent random variables to take values from different domains (see Appendix C).

Subsequently, based on the derived bound (25), we can analyze the asymptotic behavior of the learning process for domain adaptation with multiple sources.

5.2 Asymptotic Convergence

In statistical learning theory, it is well-known that the complexity of function class is one of main factors to the asymptotic convergence of the learning process under the assumption of same distribution (Vapnik, 1999; Van der Vaart and Wellner, 1996; Mendelson, 2003).

From Theorem 5.1, we can directly arrive at the following result showing that the asymptotic convergence of the learning process for domain adaptation with multiple sources is affected by the three aspects: the choice of $\mathbf{w}$, the discrepancy quantity $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and the uniform entropy number $\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi/8, 2\sum_{k=1}^K N_k\big)$.

Theorem 5.3 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a,b]$. Let $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$. If the following condition holds:*

$$\lim_{N_1,\cdots,N_K \to +\infty} \frac{\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi/8, 2\sum_{k=1}^K N_k\big)}{\frac{\big(\prod_{k=1}^K N_k\big)}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k} N_i)\big)}} < +\infty, \tag{29}$$

then we have for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$,

$$\lim_{N_1,\cdots,N_K \to +\infty} \Pr\Big\{ \sup_{f\in\mathcal{F}} \big|\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f\big| > \xi \Big\} = 0. \tag{30}$$

As shown in Theorem 5.3, if the choice of $\mathbf{w} \in [0,1]^K$ and the uniform entropy number $\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k\big)$ satisfy the condition (29) with $\sum_{k=1}^K w_k = 1$, the probability of the event that $\sup_{f\in\mathcal{F}} \big|\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f\big| > \xi$ will converge to *zero* for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, when the sample numbers $N_1, \cdots, N_K$ of multiple sources go to *infinity*, respectively. This is partially in accordance with the classical result of the asymptotic convergence of the learning process under the assumption of same distribution (*cf.* Theorem 2.3 and Definition 2.5 of [22]): the probability of the event that $\sup_{f\in\mathcal{F}} \big|\mathrm{E} f - \mathrm{E}_N f\big| > \xi$ will converge to *zero* for any $\xi > 0$, if the uniform entropy number $\ln \mathcal{N}_1(\mathcal{F}, \xi, N)$ satisfies the following:

$$\lim_{N\to+\infty} \frac{\ln \mathcal{N}_1(\mathcal{F}, \xi, N)}{N} < +\infty. \tag{31}$$

Note that in the learning process of domain adaptation with multiple sources, the uniform convergence of the empirical risk on the source domains to the expected risk on the target domain may not hold, because the limit (30) does not hold for any $\xi > 0$ but for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. By contrast, the limit (30) holds for all $\xi > 0$ in the learning process under the assumption of same distribution, if the condition (31) is satisfied. Again, these two results coincide when any source domain and the target domain match, *i.e.*, $D_{\mathcal{F}}^{(\mathbf{w})}(S,T) = D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$.

Next, we study the rate of convergence of the learning process for domain adaptation with multiple sources.

5.3 Rate of Convergence

Recalling (25), we can find that the rate of convergence is affected by the choice of $\mathbf{w}$. According to the Cauchy-Schwarz inequality, setting $w_k = N_k / \sum_{k=1}^{K} N_k$ $(1 \leq k \leq K)$, we have

$$\max\left\{\frac{\big(\prod_{k=1}^{K} N_k\big)}{32(b-a)^2\big(\sum_{k=1}^{K} w_k^2(\prod_{i\neq k} N_i)\big)}\right\} = \frac{N_1 + N_2 + \cdots + N_K}{32(b-a)^2}, \tag{32}$$

which minimizes the second term of the right-hand side of (25). Thus, by (25), (27) and (32), we find that the fastest rate of convergence of the learning process is up to $O(1/\sqrt{N})$ which is the same as the classical result (27) of the learning process under the assumption of same distribution if the discrepancy $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ is ignored.

In addition, the bound (28) based on the Rademacher complexity also implies that the rate of convergence of the learning process is affected by the choice of $\mathbf{w}$. Again, according to Cauchy-Schwarz inequality, setting $w_k = \frac{N_k}{\sum_{k=1}^{K} N_k}$ $(1 \leq k \leq K)$ leads to the fastest rate of convergence:

$$\sqrt{\frac{(b-a)^2 \ln(1/\epsilon)}{2\sum_{k=1}^{K} N_k}} = O(1/\sqrt{N}),$$

which is in accordance with the aforementioned analysis. The following numerical experiments support our theoretical findings (see Fig. 1).

5.4 Numerical Experiment

We have performed the numerical experiments to verify the theoretic analysis of the asymptotic convergence of the learning processes for domain adaptation with multiple sources. Without loss of generality, we only consider the case of $K = 2$, *i.e.,* there are two source domains and one target domain. The experiment data are generated in the following way:

For the target domain $\mathcal{Z}^{(T)} = \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^{100} \times \mathbb{R}$, we consider $\mathcal{X}^{(T)}$ as a Gaussian distribution $N(0,1)$ and draw $\{\mathbf{x}_n^{(T)}\}_{n=1}^{N_T}$ $(N_T = 4000)$ from $\mathcal{X}^{(T)}$ randomly and independently. Let $\beta \in \mathbb{R}^{100}$ be a random vector of a Gaussian distribution $N(1,5)$, and let the random vector $R \in \mathbb{R}^{100}$ be a noise term with $R \sim N(0, 0.5)$. For any $1 \leq n \leq N_T$, we randomly draw β and R from $N(1,5)$ and $N(0, 0.01)$ respectively, and then generate $y_n^{(T)} \in \mathcal{Y}$ as follows:

$$y_n^{(T)} = \langle \mathbf{x}_n^{(T)}, \beta \rangle + R. \tag{33}$$

The derived $\{(\mathbf{x}_n^{(T)}, y_n^{(T)})\}_{n=1}^{N_T}$ $(N_T = 4000)$ are the samples of the target domain $\mathcal{Z}^{(T)}$ and will be used as the test data.

In the similar way, we generate the sample set $\{(\mathbf{x}_n^{(1)}, y_n^{(1)})\}_{n=1}^{N_1}$ $(N_1 = 2000)$ of the source domain $\mathcal{Z}^{(S_1)} = \mathcal{X}^{(1)} \times \mathcal{Y}^{(1)} \subset \mathbb{R}^{100} \times \mathbb{R}$: for any $1 \leq n \leq N_1$,

$$y_n^{(1)} = \langle \mathbf{x}_n^{(1)}, \beta \rangle + R, \tag{34}$$

where $\mathbf{x}_n^{(1)} \sim N(0.5, 1)$, $\beta \sim N(1,5)$ and $R \sim N(0, 0.5)$.

For the source domain $\mathcal{Z}^{(S_2)} = \mathcal{X}^{(2)} \times \mathcal{Y}^{(2)} \subset \mathbb{R}^{100} \times \mathbb{R}$, the samples $\{(\mathbf{x}_n^{(2)}, y_n^{(2)})\}_{n=1}^{N_2}$ $(N_2 = 2000)$ are derived in the following way: for any $1 \leq n \leq N_2$,

$$y_n^{(2)} = \langle \mathbf{x}_n^{(2)}, \beta \rangle + R, \tag{35}$$

where $\mathbf{x}_n^{(2)} \sim N(2,5)$, $\beta \sim N(1,5)$ and $R \sim N(0,0.5)$.

In this experiment, we use the method of Least Square Regression[1] to minimize the empirical risk

$$\mathrm{E}_w^{(S)}(\ell \circ g) = \frac{w}{N_1}\sum_{n=1}^{N_1} \ell(g(\mathbf{x}_n^{(1)}), y_n^{(1)}) + \frac{(1-w)}{N_2}\sum_{n=1}^{N_2} \ell(g(\mathbf{x}_n^{(2)}), y_n^{(2)}) \tag{36}$$

for different combination coefficients $w \in \{0.1, 0.3, 0.5, 0.9\}$ and then compute the discrepancy $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ for each $N_1 + N_2$. The initial N_1 and N_2 both equal to 200. Each test is repeated 30 times and the final result is the average of the 30 results. After each test, we increment N_1 and N_2 by 200 until $N_1 = N_2 = 2000$. The experiment results are shown in Fig. 1.

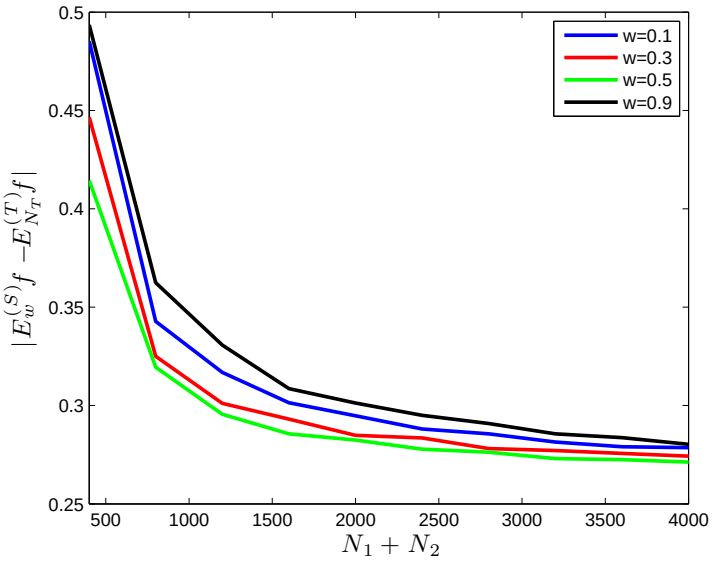

Figure 1: Domain Adaptation with Multiple Sources

From Fig. 1, we can find that for any choice of w, the curve of $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ is decreasing when $N_1 + N_2$ increases, which is in accordance with the results presented in Theorems 5.1 & 5.3. Moreover, when $w = 0.5$, the discrepancy $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ has the fastest rate of convergence, and the rate becomes slower as w is further away from 0.5. In this experiment, we set $N_1 = N_2$ that implies that $N_2/(N_1 + N_2) = 0.5$. Recalling (25), we have shown that $w = N_2/(N_1 + N_2)$ will provide the fastest rate of convergence and this proposition is supported by the experiment results shown in Fig. 1.

6 Learning Process of Domain Adaptation Combining Source and Target Data

In this section, we present two generalization bounds of the learning process for domain adaptation combining source and target data, which are based on the uniform entropy number and

[1]SLEP Package: http://www.public.asu.edu/~jye02/Software/SLEP/index.htm

the Rademacher complexity, respectively. We then analyze the asymptotic convergence and the rate of convergence of the learning process in addition to the numerical experiments supporting our theoretical analysis.

6.1 Generalization Bounds

The following theorem provides a generalization bound based on the uniform entropy number with respect to the metric ℓ_1^τ defined in (22). Similar to the situation of domain adaptation with multiple sources, the proof of this theorem is achieved by using a specific Hoeffding-type deviation inequality and a symmetrization inequality for domain adaptation combining source and target data (see Appendix B).

Theorem 6.1 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a,b]$. Let $\mathbf{Z}_1^{N_S} = \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} = \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be two sets of i.i.d. samples drawn from domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Then, for any $\tau \in [0,1)$ and given an arbitrary $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$, with probability at least $1-\epsilon$,*

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_\tau f - \mathrm{E}^{(T)} f\right| \leq (1-\tau)D_{\mathcal{F}}(S,T) + \left(\frac{\ln \mathcal{N}_1^\tau(\mathcal{F}, \xi'/8, 2(N_S+N_T)) - \ln(\epsilon/8)}{\frac{N_S N_T}{32(b-a)^2((1-\tau)^2 N_T + \tau^2 N_S)}}\right)^{\frac{1}{2}}, \tag{37}$$

where $D_{\mathcal{F}}(S,T)$ is defined in (12) and $\xi' := \xi - (1-\tau)D_{\mathcal{F}}(S,T)$.

Compared to the classical result (27) under the assumption of same distribution, the derived bound (37) contains a term of discrepancy quantity $(1-\tau)D_{\mathcal{F}}(S,T)$ that is determined by two factors: the combination coefficient τ and the quantity $D_{\mathcal{F}}(S,T)$. The two results coincide when the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.,* $D_{\mathcal{F}}(S,T) = 0$.

Based on the Rademacher complexity, we then get another generalization bound of the learning process for domain adaptation combining source and target data. Its proof is postponed in Appendix C.

Theorem 6.2 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a,b]$. Let $\mathbf{Z}_1^{N_S} = \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} = \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be two sets of i.i.d. samples drawn from the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Then, given $\tau \in [0,1)$ and for any $\epsilon > 0$, we have with probability at least $1-\epsilon$,*

$$\begin{aligned}\sup_{f\in\mathcal{F}} \left|\mathrm{E}_\tau f - \mathrm{E}^{(T)} f\right| \leq & (1-\tau)D_{\mathcal{F}}(S,T) + 2(1-\tau)\mathcal{R}^{(S)}(\mathcal{F}) \\ & + 2\tau\mathcal{R}_{N_T}^{(T)}(\mathcal{F}) + 3\tau\sqrt{\frac{(b-a)\ln(4/\epsilon)}{2N_T}} \\ & + (1-\tau)\sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)},\end{aligned} \tag{38}$$

where $D_{\mathcal{F}}(S,T)$ is defined in (12).

Note that in the derived bound (38), we adopt an empirical Rademacher complexity $\mathcal{R}_{N_T}^{(T)}(\mathcal{F})$ that is based on the data drawn from the target domain $\mathcal{Z}^{(T)}$, because the distribution of $\mathcal{Z}^{(T)}$ is unknown in the situation of domain adaptation. Similar to the aforementioned discussion, the generalization bound (38) coincides with the result under the assumption of same distribution (see Bousquet *et al.*, 2004, Theorem 5), when the source domain of $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.*, $D_{\mathcal{F}}(S,T)=0$.

The two results (37) and (38) exhibit a tradeoff between the sample numbers N_S and N_T, which is associated with the choice of τ. Although the tradeoff has been mentioned in some previous works (see Blitzer *et al.*, 2008; Ben-David *et al.*, 2010), the following will show a rigorous theoretical analysis of the tradeoff.

6.2 Asymptotic Convergence

Following Theorem 6.1, we can directly obtain the concerning result pointing out that the asymptotic convergence of the learning process for domain adaptation combining source and target data is affected by three factors: the uniform entropy number $\ln \mathcal{N}_1^{\tau}(\mathcal{F}, \xi/8, 2(N_S+N_T))$, the integral probability metric $D_{\mathcal{F}}(S,T)$ and the choice of $\tau \in [0,1)$.

Theorem 6.3 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a,b]$. Given a $\tau \in [0,1)$, if the following condition holds:*

$$\lim_{N_S \to +\infty} \frac{\ln \mathcal{N}_1^{\tau}(\mathcal{F}, \xi'/8, 2(N_S+N_T))}{\frac{N_S N_T}{((1-\tau)^2 N_T + \tau^2 N_S)}} < +\infty \tag{39}$$

with $\xi' := \xi - (1-\tau) D_{\mathcal{F}}(S,T)$, then we have for any $\xi > (1-\tau) D_{\mathcal{F}}(S,T)$,

$$\lim_{N_S \to +\infty} \Pr\left\{ \sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_{\tau} f \right| > \xi \right\} = 0. \tag{40}$$

As shown in Theorem 6.3, if the choice of $\tau \in [0,1)$ and the uniform entropy number $\ln \mathcal{N}_1^{\tau}(\mathcal{F}, \xi'/8, 2(N_S+N_T))$ satisfy the condition (39), the probability of the event $\sup_{f \in \mathcal{F}} \left| \mathrm{E}^{(T)} f - \mathrm{E}_{\tau} f \right| > \xi$ will converge to *zero* for any $\xi > (1-\tau) D_{\mathcal{F}}(S,T)$, when N_S goes to *infinity*. This is partially in accordance with the classical result under the assumption of same distributions derived from the combination of Theorem 2.3 and Definition 2.5 of Mendelson (2003).

Note that in the learning process for domain adaptation combining source and target data, the uniform convergence of the empirical risk $\mathrm{E}_{\tau} f$ to the expected risk $\mathrm{E}^{(T)} f$ may not hold, because the limit (40) does not hold for any $\xi > 0$ but for any $\xi > (1-\tau) D_{\mathcal{F}}(S,T)$. By contrast, the limit (40) holds for all $\xi > 0$ in the learning process under the assumption of same distribution, if the condition (31) is satisfied. The two results coincide when the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.*, $D_{\mathcal{F}}(S,T)=0$.

6.3 Rate of Convergence

We consider the choice of τ that is an essential factor to the rate of convergence for the learning process and is associated with the tradeoff between the sample numbers N_S and N_T. Recalling

(37), if we fix the value of $\ln \mathcal{N}_1^{\tau}(\mathcal{F}, \xi'/8, 2(N_S+N_T))$, setting $\tau = \frac{N_T}{N_T+N_S}$ minimizes the second term of the right-hand side of (37) and then we arrive at

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_{\tau} f - \mathrm{E}^{(T)} f\right| \leq \frac{N_S D_{\mathcal{F}}(S,T)}{N_S+N_T} + \left(\frac{(\ln \mathcal{N}_1^{\tau}(\mathcal{F}, \xi'/8, 2(N_S+N_T)) - \ln(\epsilon/8))}{\frac{N_S+N_T}{32(b-a)^2}} \right)^{\frac{1}{2}}, \tag{41}$$

which implies that setting $\tau = \frac{N_T}{N_T+N_S}$ can result in the fastest rate of convergence, while it can also cause the relatively larger discrepancy between the empirical risk $\mathrm{E}_{\tau} f$ and the expected risk $\mathrm{E}^{(T)} f$, because the situation of domain adaptation is set up in the condition that $N_T \ll N_S$, which implies that $\frac{N_S}{N_S+N_T} \approx 1$. Moreover, this choice of τ associated with a trade off between sample numbers N_S and N_T is also suitable to the Rademacher-complexity-based bound (38). It is noteworthy that the value $\tau = \frac{N_T}{N_T+N_S}$ has been mentioned in the section of "Experimental Results" in Blitzer *et al.* (2008). Here, we show a rigorous theoretical analysis of this value and the following numerical experiment also supports this finding (see Fig. 2).

6.4 Numerical Experiments

In the situation of domain adaptation combining source and target data, the samples $\{(\mathbf{x}_n^{(T)}, y_n^{(T)})\}_{n=1}^{N_T}$ ($N_T = 4000$) of the target domain $\mathcal{Z}^{(T)}$ are generated in the aforementioned way (see (33)). We randomly pick $N_T' = 100$ samples from them to form the objective function and the rest $N_T'' = 3900$ are used to test.

In the similar way, the samples $\{(\mathbf{x}_n^{(S)}, y_n^{(S)})\}_{n=1}^{N_S}$ ($N_S = 4000$) of the source domain $\mathcal{Z}^{(S)}$ are generated as follows: for any $1 \leq n \leq N_S$,

$$y_n^{(S)} = \langle \mathbf{x}_n^{(S)}, \beta \rangle + R, \tag{42}$$

where $\mathbf{x}_n^{(S)} \sim N(1,2)$, $\beta \sim N(1,5)$ and $R \sim N(0, 0.5)$.

We also use the method of Least Square Regression to minimize the empirical risk

$$\mathrm{E}_{\tau}(\ell \circ g) = \frac{1-\tau}{N_S} \sum_{n=1}^{N_S} \ell(g(\mathbf{x}_n^{(1)}), y_n^{(1)}) + \frac{\tau}{N_T'} \sum_{n=1}^{N_T'} \ell(g(\mathbf{x}_n^{(T)}), y_n^{(T)})$$

for different combination coefficients $\tau \in \{0.1, 0.3, 0.5, 0.9\}$ and then compute the discrepancy $|E_{\tau} f - E_{N_T''}^{(T)} f|$ for each N_S. Since it has to be satisfied that $N_S \gg N_T'$, the initial N_S is set to be 200. Each test is repeated 100 times and the final result is the average of the 100 results. After each test, we increment N_S by 200 until $N_S = 4000$. The experiment results are shown in Fig. 2.

Figure (2) illustrates that for any choice of $\tau \in \{0.1, 0.3, 0.5, 0.9\}$, the curve of $|E_{\tau} f - E_{N_T''}^{(T)}|$ is decreasing as N_S increases. This is in accordance with our results of the asymptotic convergence of the learning process for domain adaptation with multiple sources (see Theorems 6.1 and 6.3). Furthermore, Fig. 2 also shows that when $\tau \approx N_T'/(N_S+N_T')$, the discrepancy $|E_{\tau}^{(S)} f - E_{N_T''}^{(T)} f|$ has the fastest rate of convergence, and the rate becomes slower as τ is further away from $N_T'/(N_S+N_T')$. Thus, this is in accordance with the theoretical analysis of the asymptotic convergence presented above.

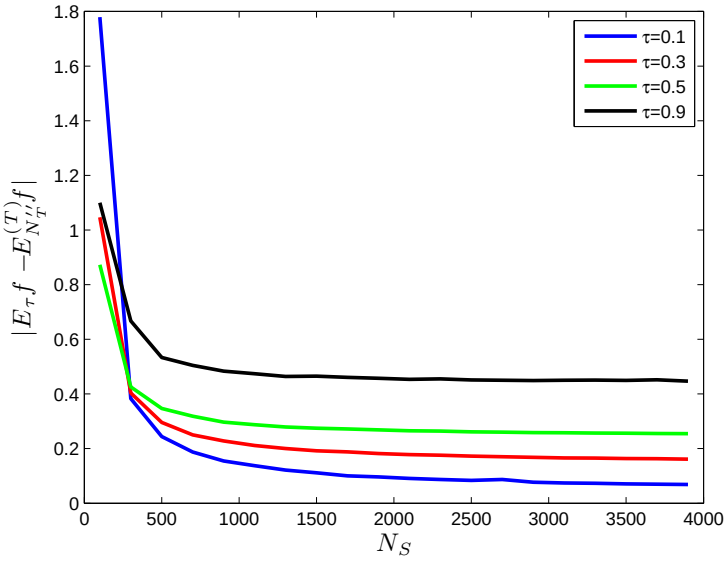

Figure 2: Domain Adaptation Combining Source and Target Data

7 Prior Works

There have been some previous works on the theoretical analysis of domain adaptation with multiple sources (see Ben-David *et al.*, 2010; Crammer *et al.*, 2006, 2008; Mansour *et al.*, 2008, 2009a) and domain adaptation combining source and target data (see Blitzer *et al.*, 2008; Ben-David *et al.*, 2010).

In Crammer *et al.* (2006, 2008), the function class and the loss function are assumed to satisfy the conditions of "α-triangle inequality" and "uniform convergence bound". Moreover, one has to get some prior information about the disparity between any source domain and the target domain. Under these conditions, some generalization bounds were obtained by using the classical techniques developed under the assumption of same distribution.

Mansour *et al.* (2008) proposed another framework to study the problem of domain adaptation with multiple sources. In this framework, one has to know some prior knowledge including the exact distributions of the source domains and the hypothesis function with a small loss on each source domain. Furthermore, the target domain and the hypothesis function on the target domain were deemed as the mixture of the source domains and the mixture of the hypothesis functions on the source domains, respectively. Then, by introducing the Rényi divergence, Mansour *et al.* (2009a) extended their previous work (Mansour *et al.*, 2008) to a more general setting, where the distribution of the target domain can be arbitrary and one only needs to know an approximation of the exact distribution of each source domain. Ben-David *et al.* (2010) also discussed the situation of domain adaptation with the mixture of source domains.

In Ben-David *et al.* (2010); Blitzer *et al.* (2008), domain adaptation combining source and target data was originally proposed and meanwhile a theoretical framework was presented to analyze its properties for the classification tasks by introducing the $\mathcal{H}$-divergence. Under the condition of "λ-close", the authors applied the classical techniques developed under the assumption of same distribution to achieve the generalization bounds based on the VC dimension.

Mansour *et al.* (2009b) introduced the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ to capture the difference between domains and this quantity can be used in both classification and regression tasks. By applying the classical results of statistical learning theory, the authors obtained the generalization bounds based on the Rademacher complexity.

8 Conclusion

In this paper, we propose a new framework to obtain generalization bounds of the learning process for two representative types of domain adaptation: domain adaptation with multiple sources and domain adaptation combining source and target data. This framework is suitable for a variant of learning tasks including classification and regression. Based on the derived bounds, we theoretically analyze the asymptotic convergence and the rate of convergence of the learning process for domain adaptation. There are four important aspects of this framework: the quantity measuring the difference between two domains; the complexity measure of function class, the deviation inequality and the symmetrization inequality for domain adaptation.

- We use the integral probability metric $D_{\mathcal{F}}(S, T)$ to measure the difference between two domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$. We show that the integral probability metric is well-defined and is a (semi)metric on the space of the probability distributions. It can be bounded by the summation of the *discrepancy distance* $\mathrm{disc}_\ell(\mathcal{D}^{(S)}, \mathcal{D}^{(T)})$ and the quantity $Q_{\mathcal{G}}^{(T)}(g_*^{(S)}, g_*^{(T)})$, which measure the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between labeling functions $g_*^{(S)}$ and $g_*^{(T)}$, respectively. Note that there is a special case that is more suitable to the integral probability metric $D_{\mathcal{F}}(S, T)$ than other quantities (see Remark 3.1).

- The uniform entropy number and the Rademacher complexity are adopted to achieved the generalization bounds (25); (37) and (28); (38), respectively. It is noteworthy that the generalization bounds (25) and (37) can lead to the results based on the fat-shattering dimension, respectively (see Mendelson, 2003, Theorem 2.18). According to Theorem 2.6.4 of Van der Vaart and Wellner (1996), the bounds based on the VC dimension can also be obtained from the results (25) and (37), respectively.

- Instead of directly applying the classical techniques, we present the specific deviation inequalities for the learning process of domain adaptation. In order to obtain the generalization bounds based on the uniform entropy numbers, we develop the specific Hoeffding-type deviation inequalities for the two types of domain adaptation, respectively (see Appendices A & B). Furthermore, we also generalize the classical McDiarmid's inequality to a more general setting where the independent random variables can take value from different domains (see Appendix C).

- We also develop the related symmetrization inequalities of the learning process for domain adaptation. The derived inequalities incorporate the discrepancy term that is determined by the difference between the source and the target domains and reflects the learning-transfering from the source to the target domains.

Based on the derived generalization bounds, we provide a rigorous theoretical analysis of the asymptotic convergence and the rate of convergence of the learning process for either kind of

domain adaptation. We also consider the choices of $\mathbf{w}$ and τ that affect the rate of convergence of the learning processes for the two types of domain adaptation, respectively. Moreover, we give a comparison with the previous works Ben-David *et al.* (2010); Crammer *et al.* (2006, 2008); Mansour *et al.* (2008, 2009a); Blitzer *et al.* (2008) as well as the related results of the learning process under the assumption of same distribution (see Bousquet *et al.*, 2004; Mendelson, 2003). The numerical experiments support our theoretical findings as well.

In our future work, we will attempt to find a new distance between distributions to develop the generalization bounds based on other complexity measures, and analyze other theoretical properties of domain adaptation.

References

V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, New York, 1999.

O. Bousquet, S. Boucheron and G. Lugosi. Introduction to Statistical Learning Theory. *Lecture Notes in Artificial Intelligence*, 3176:169-207, 2004.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics (Springer Series in Statistics)*. Springer, 1996.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM*, 36(4):929-965, 1989.

P.L. Bartlett, O. Bousquet and S. Mendelson. Local Rademacher Complexities. *Annals of Statistics*, 33:1497-1537, 2005.

Z. Hussain and J. Shawe-Taylor. Improved Loss Bounds for Multiple Kernel Learning. *Journal of Machine Learning Research*, **15(W&CP)**:370-377, 2011.

J. Jiang, and C. Zhai. Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 264-271, 2007.

J. Blitzer, M. Dredze and F. Pereira. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 440-447, 2007.

S. Bickel, M. Brückner and T. Scheffer. Discriminative learning for differing training and test distributions. *Proceedings of the 24th international conference on Machine learning (ICML)*, 81-88, 2007.

P. Wu and T.G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. *Proceedings of the twenty-first international conference on Machine learning,* 2004.

J. Blitzer, R. McDonald and F. Pereira. Domain adaptation with structural correspondence learning. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira and J. Wortman. A Theory of Learning from Different Domains. *Machine Learning*, 79:151-175, 2010.

W. Bian, D. Tao and Y. Rui. Cross-Domain Human Action Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):298-307, 2012.

K. Crammer, M. Kearns and J. Wortman. Learning from Multiple Sources. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

K. Crammer, M. Kearns and J. Wortman. Learning from Multiple Sources. *Journal of Machine Learning Research*, 9:1757-1774, 2008.

Y. Mansour, M. Mohri and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Y. Mansour, M. Mohri and A. Rostamizadeh. Multiple Source Adaptation and The Rényi Divergence. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira and J. Wortman. Learning Bounds for Domain Adaptation. *Advances in Neural Information Processing Systems (NIPS)*, 2008.

Y. Mansour, M. Mohri and A. Rostamizadeh. Domain Adaptation: Learning Bounds and Algorithms. *Conference on Learning Theory (COLT)*, 2009.

S. Ben-David, J. Blitzer, K. Crammer and F. Pereira. Analysis of Representations for Domain Adaptation. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

V.M. Zolotarev. Probability Metrics. *Theory of Probability and its Application*, 28(1):278-302, 1984.

S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models.* John Wiley and Sons, 1991.

A. Müller. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429-443, 1997.

M.D. Reid and R.C. Williamson. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research,* 12:731-817, 2011.

B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G.R.G. Lanckriet and B. Schölkopf. A Note on Integral Probability Metrics and ϕ-Divergences. *CoRR*, abs/0901.2698, 2009.

B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf and G.R.G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550-1599, 2012.

S. Mendelson. A Few Notes on Statistical Learning Theory. *Advanced Lectures on Machine Learning*, 2600:1-40, 2003.

W. Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13-30, 1963.

A Proof of Theorem 5.1

In this appendix, we provide the proof of Theorem 5.1. In order to achieve the proof, we need to develop the specific Hoeffding-type deviation inequality and the symmetrization inequality for domain adaptation with multiple sources.

A.1 Hoeffding-Type Deviation Inequality for Multiple Sources

Deviation (or concentration) inequalities play an essential role in obtaining the generalization bounds for a certain learning process. Generally, specific deviation inequalities need to be developed for different learning processes. There are many popular deviation and concentration inequalities, *e.g.*, Hoeffding's inequality, McDiarmid's inequality, Bennett's inequality, Bernstein's inequality and Talagrand's inequality. These results are all built under the assumption of same distribution, and thus they are not applicable (or at least cannot be directly applied) to the setting of multiple sources. Next, based on Hoeffding's inequality (Hoeffding, 1963), we present a deviation inequality for multiple sources.

Theorem A.1 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a,b]$. Let $\mathbf{Z}_1^{N_k}=\{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}$ be the set of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S_k)}\subset\mathbb{R}^L$ $(1\leq k\leq K)$. Given $\mathbf{w}\in[0,1]^K$ with $\sum_{k=1}^K w_k=1$ and for any $f\in\mathcal{F}$, we define a function $F_{\mathbf{w}}:\mathbb{R}^{L\sum_{k=1}^K N_k}\to\mathbb{R}$ as*

$$F_{\mathbf{w}}\Big(\{\mathbf{X}_1^{N_k}\}_{k=1}^K\Big):=\sum_{k=1}^K w_k\Big(\prod_{i\neq k}N_i\Big)\sum_{n=1}^{N_k}f(\mathbf{x}_n^{(k)}),\tag{43}$$

where for any $1\leq k\leq K$ and given $N_k\in\mathbb{N}$, the set $\mathbf{X}_1^{N_k}$ is denoted as

$$\mathbf{X}_1^{N_k}:=\{\mathbf{x}_1^{(k)},\mathbf{x}_2^{(k)},\cdots,\mathbf{x}_{N_k}^{(k)}\}\in(\mathbb{R}^L)^{N_k}.$$

Then, we have for any $\xi>0$,

$$\begin{aligned}&\Pr\Big\{\big|\mathrm{E}^{(S)}F_{\mathbf{w}}-F_{\mathbf{w}}\big(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\big)\big|>\xi\Big\}\\ \leq&2\exp\left\{-\frac{2\xi^2}{(b-a)^2\big(\prod_{k=1}^K N_k\big)\Big(\sum_{k=1}^K w_k^2\big(\prod_{i\neq k}N_i\big)\Big)}\right\},\end{aligned}\tag{44}$$

where $\mathrm{E}^{(S)}$ stands for the expectation taken on all source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$.

This result is an extension of the classical Hoeffding-type deviation inequality under the assumption of same distribution (see Bousquet *et al.*, 2004, Theorem 1). Compared to the classical result, the resultant deviation inequality (44) is suitable to the setting of multiple sources. These two inequalities coincide when there is only one source, *i.e.*, $K=1$

The proof of Theorem A.1 is processed by a martingale method. Before the formal proof, we introduce some essential notations.

Let $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ be sample sets drawn from multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$, respectively. Define a random variable

$$S_n^{(k)}:=\mathrm{E}^{(S)}\Big\{F_{\mathbf{w}}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)|\mathbf{Z}_1^{N_1},\mathbf{Z}_1^{N_2},\cdots,\mathbf{Z}_1^{N_{k-1}},\mathbf{Z}_1^{n}\Big\},\ \ 1\leq k\leq K,\ 0\leq n\leq N_k,\tag{45}$$

where

$$\mathbf{Z}_1^n = \{\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \cdots, \mathbf{z}_n^{(k)}\} \subseteq \mathbf{Z}_1^{N_k}, \text{ and } \mathbf{Z}_1^0 = \varnothing.$$

It is clear that

$$S_0^{(1)} = \mathrm{E}^{(S)} F_{\mathbf{w}} \text{ and } S_{N_K}^{(K)} = F_{\mathbf{w}}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K),$$

where $\mathrm{E}^{(S)}$ stands for the expectation taken on all source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$.

Then, according to (43) and (45), we have for any $1 \le k \le K$ and $1 \le n \le N_k$,

$$\begin{aligned}
S_n^{(k)} - S_{n-1}^{(k)} =& \mathrm{E}^{(S)} \left\{ F_{\mathbf{w}}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K) \middle| \mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n \right\} \\
& - \mathrm{E}^{(S)} \left\{ F_{\mathbf{w}}(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K) \middle| \mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1} \right\} \\
=& \mathrm{E}^{(S)} \left\{ \sum_{k=1}^K w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)}) \middle| \mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n \right\} \\
& - \mathrm{E}^{(S)} \left\{ \sum_{k=1}^K w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)}) \middle| \mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1} \right\} \\
=& \sum_{l=1}^{k-1} w_l \Big(\prod_{i \neq l} N_i \Big) \sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{j=1}^{n} f(\mathbf{z}_j^{(k)}) \\
& + \mathrm{E}^{(S)} \left\{ \sum_{l=k+1}^K w_l \Big(\prod_{i \neq l} N_i \Big) \sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{j=n+1}^{N_k} f(\mathbf{z}_j^{(k)}) \right\} \\
& - \sum_{l=1}^{k-1} w_l \Big(\prod_{i \neq l} N_i \Big) \sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) - w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{j=1}^{n-1} f(\mathbf{z}_j^{(k)}) \\
& - \mathrm{E}^{(S)} \left\{ \sum_{l=k+1}^K w_l \Big(\prod_{i \neq l} N_i \Big) \sum_{j=1}^{N_l} f(\mathbf{z}_j^{(l)}) + w_k \Big(\prod_{i \neq k} N_i \Big) \sum_{j=n}^{N_k} f(\mathbf{z}_j^{(k)}) \right\} \\
=& w_k \Big(\prod_{i \neq k} N_i \Big) \left(f(\mathbf{z}_n^{(k)}) - \mathrm{E}^{(S_k)} f \right). \qquad (46)
\end{aligned}$$

To prove Theorem A.1, we need the following inequality resulted from Hoeffding's lemma.

Lemma A.1 *Let f be a function with the range $[a, b]$. Then, the following holds for any $\alpha > 0$:*

$$\mathrm{E}\left\{ \mathrm{e}^{\alpha(f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)} f)} \right\} \le \mathrm{e}^{\frac{\alpha^2 (b-a)^2}{8}}. \qquad (47)$$

Proof. We consider

$$(f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)} f)$$

as a random variable. Then, it is clear that

$$\mathrm{E}\{f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)} f\} = 0.$$

Since the value of $\mathrm{E}^{(S)} f$ is a constant denoted as e, we have

$$a - e \le f(\mathbf{z}^{(S)}) - \mathrm{E}^{(S)} f \le b - e.$$

According to Hoeffding's lemma, we then have

$$\mathrm{E}\left\{\mathrm{e}^{\alpha(f(\mathbf{z}^{(S)})-\mathrm{E}^{(S)}f)}\right\} \leq \mathrm{e}^{\frac{\alpha^2(b-a)^2}{8}}. \tag{48}$$

This completes the proof. ■

We are now ready to prove Theorem A.1.

Proof of Theorem A.1. According to (43), (46), Lemma A.1, Markov's inequality, Jensen's inequality and the law of iterated expectation, we have for any $\alpha > 0$,

$$\begin{aligned}
&\Pr\left\{F_{\mathbf{w}}\big(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\big) - \mathrm{E}^{(S)}F_{\mathbf{w}} > \xi\right\}\\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right)-\mathrm{E}^{(S)}F_{\mathbf{w}}\right)}\right\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{E}\left\{\mathrm{e}^{\alpha\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)}\big|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{K-1}},\mathbf{Z}_1^{N_K-1}\right\}\right\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)}\big|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{K-1}},\mathbf{Z}_1^{N_K-1}\right\}\right\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\mathrm{E}\left\{\mathrm{e}^{\alpha w_K(\prod_{i\neq K}N_i)(f(\mathbf{z}_N^{(K)})-\mathrm{E}^{(S_K)}f)}\right\}\right\}\\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{k=1}^K\sum_{n=1}^{N_k}\left(S_n^{(k)}-S_{n-1}^{(k)}\right)-\left(S_{N_K}^{(K)}-S_{N_K-1}^{(K)}\right)\right)}\right\}\mathrm{e}^{\frac{\alpha^2w_K^2(\prod_{i\neq K}N_i)^2(b-a)^2}{8}},
\end{aligned} \tag{49}$$

where $\mathbf{Z}_1^{N_K-1} := \{\mathbf{z}_1^{(K)},\cdots,\mathbf{z}_{N_K-1}^{(K)}\} \subset \mathbf{Z}_1^{N_K}$. Therefore, we have

$$\Pr\left\{F_{\mathbf{w}}\left(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\right) - \mathrm{E}^{(S)}F_{\mathbf{w}} > \xi\right\} \leq \mathrm{e}^{\Phi(\alpha)-\alpha\xi}, \tag{50}$$

where

$$\Phi(\alpha) = \frac{\alpha^2(b-a)^2\big(\prod_{k=1}^K N_k\big)\Big(\sum_{k=1}^K w_k^2\big(\prod_{i\neq k}N_i\big)\Big)}{8}. \tag{51}$$

Similarly, we can obtain

$$\Pr\left\{\mathrm{E}^{(S)}F_{\mathbf{w}} - F_{\mathbf{w}}\big(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\big) > \xi\right\} \leq \mathrm{e}^{\Phi(\alpha)-\alpha\xi}. \tag{52}$$

Note that $\Phi(\alpha)-\alpha\xi$ is a quadratic function with respect to $\alpha > 0$ and thus the minimum value "$\min_{\alpha>0}\{\Phi(\alpha)-\alpha\xi\}$" is achieved when

$$\alpha = \frac{4\xi}{(b-a)^2\big(\prod_{k=1}^K N_k\big)\Big(\sum_{k=1}^K w_k^2\big(\prod_{i\neq k}N_i\big)\Big)}.$$

By combining (50), (51) and (52), we arrive at

$$\begin{aligned}
&\Pr\left\{\left|\mathrm{E}^{(S)}F_{\mathbf{w}} - F_{\mathbf{w}}\big(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K\big)\right| > \xi\right\}\\
\leq&2\exp\left\{-\frac{2\xi^2}{(b-a)^2\big(\prod_{k=1}^K N_k\big)\Big(\sum_{k=1}^K w_k^2\big(\prod_{i\neq k}N_i\big)\Big)}\right\}.
\end{aligned}$$

This completes the proof. ■

In the following subsection, we present a symmetrization inequality for domain adaptation with multiple sources.

A.2 Symmetrization Inequality

Symmetrization inequalities are mainly used to replace the expected risk by an empirical risk computed on another sample set that is independent of the given sample set but has the same distribution. In this manner, the generalization bounds can be achieved by applying some kinds of complexity measures, *e.g.*, the covering number and the VC dimension. However, the classical symmetrization results are built under the assumption of same distribution (see Bousquet *et al.*, 2004). The symmetrization inequality for domain adaptation with multiple sources is presented in the following theorem:

Theorem A.2 *Assume that $\mathcal{F}$ is a function class with the range $[a, b]$. Let sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ and $\{\mathbf{Z'}_1^{N_k}\}_{k=1}^K$ be drawn from the source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$. Then, given an arbitrary $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, we have for any $\left(\prod_{k=1}^K N_k\right) \geq \frac{8(b-a)^2}{(\xi')^2}$,*

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \xi\right\} \leq 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}'^{(S)}_{\mathbf{w}}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \frac{\xi'}{2}\right\}, \tag{53}$$

where $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$.

This theorem shows that given $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \xi$$

can be bounded by using the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}'^{(S)}_{\mathbf{w}}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right| > \frac{\xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)}{2} \tag{54}$$

that is only determined by the characteristics of the source domains $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ when $\prod_{k=1}^K N_k \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. Compared to the classical symmetrization result under the assumption of same distribution (see Bousquet *et al.*, 2004), there is a discrepancy term $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ in the derived inequality. Especially, the two results coincide when any source domain and the target domain match, *i.e.*, $D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$. The following is the proof of Theorem A.2.

Proof of Theorem A.2. Let $\widehat{f}$ be the function achieving the supremum:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_{\mathbf{w}}^{(S)}f\right|$$

with respect to the sample set $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$. According to (6), (7), (12) and (26), we arrive at

$$|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| = |\mathrm{E}^{(T)}\widehat{f} - \overline{\mathrm{E}}^{(S)}\widehat{f} + \overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right|, \tag{55}$$

and thus,

$$\Pr\left\{|\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\xi\right\}\leq\Pr\left\{D_{\mathcal{F}}^{(\mathbf{w})}(S,T)+\left|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}\right|>\xi\right\},\tag{56}$$

where the expectation $\mathrm{E}^{(S)}\widehat{f}$ is defined as

$$\overline{\mathrm{E}}^{(S)}\widehat{f}:=\sum_{k=1}^{K}w_k\mathrm{E}^{(S_k)}\widehat{f}.\tag{57}$$

Let

$$\xi':=\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T),\tag{58}$$

and denote $\wedge$ as the conjunction of two events. According to the triangle inequality, we have

$$\left(|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|-|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}-\overline{\mathrm{E}}^{(S)}\widehat{f}|\right)\leq|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|,$$

and thus for any $\xi'>0$,

$$\begin{aligned}&\left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\xi'}\right)\left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}|<\frac{\xi'}{2}}\right)\\=&\mathbf{1}_{\left\{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\xi'\right\}\wedge\left\{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}|<\frac{\xi'}{2}\right\}}\\\leq&\mathbf{1}_{|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\frac{\xi'}{2}}.\end{aligned}$$

Then, taking the expectation with respect to $\{\mathbf{Z}'^{N_k}_1\}_{k=1}^K$ gives

$$\begin{aligned}&\left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\xi'}\right)\Pr'\left\{|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}|<\frac{\xi'}{2}\right\}\\\leq&\Pr'\left\{|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}-\mathrm{E}_{\mathbf{w}}^{(S)}\widehat{f}|>\frac{\xi'}{2}\right\}.\end{aligned}\tag{59}$$

By Chebyshev's inequality, since $\{\mathbf{Z}'^{N_k}_1\}_{k=1}^K$ are the sets of i.i.d. samples drawn from the multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ respectively, we have for any $\xi'>0$,

$$\begin{aligned}\Pr'\left\{\left|\overline{\mathrm{E}}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f}\right|\geq\frac{\xi'}{2}\right\}\leq&\Pr'\left\{\sum_{k=1}^{K}\frac{w_k}{N_k}\sum_{n=1}^{N_k}|\mathrm{E}^{(S_k)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(k)}_n)|\geq\frac{\xi'}{2}\right\}\\=&\Pr'\left\{\sum_{k=1}^{K}w_k\Big(\prod_{i\neq k}N_i\Big)\sum_{n=1}^{N_k}|\mathrm{E}^{(S_k)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(k)}_n)|\geq\frac{\xi'\prod_{k=1}^{K}N_k}{2}\right\}\\\leq&\frac{4\mathrm{E}\left\{\sum_{k=1}^{K}w_k\big(\prod_{i\neq k}N_i\big)\sum_{n=1}^{N_k}\left|\mathrm{E}^{(S_k)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(k)}_n)\right|^2\right\}}{\big(\prod_{k=1}^{K}N_k\big)^2(\xi')^2}\\=&\frac{4\mathrm{E}\left\{\sum_{k=1}^{K}w_k\big(\prod_{i\neq k}N_i\big)N_k\,(b-a)^2\right\}}{\big(\prod_{k=1}^{K}N_k\big)^2(\xi')^2}\\=&\frac{4\big(\prod_{k=1}^{K}N_k\big)\,(b-a)^2}{\big(\prod_{k=1}^{K}N_k\big)^2(\xi')^2}=\frac{4\,(b-a)^2}{(\xi')^2\big(\prod_{k=1}^{K}N_k\big)}.\end{aligned}\tag{60}$$

Subsequently, according to (59) and (60), we have for any $\xi' > 0$,

$$\Pr'\left\{|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f} - \mathrm{E}^{(S)}_{\mathbf{w}}\widehat{f}| > \frac{\xi'}{2}\right\} \geq \left(\mathbf{1}_{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{\mathbf{w}}\widehat{f}| > \xi'}\right)\left(1 - \frac{4\,(b-a)^2}{(\xi')^2\big(\prod_{k=1}^{K} N_k\big)}\right). \tag{61}$$

By combining (56), (58) and (61), taking the expectation with respect to $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ and letting

$$\frac{4\,(b-a)^2}{(\xi')^2\big(\prod_{k=1}^{K} N_k\big)} \leq \frac{1}{2}$$

can lead to: for any $\xi > D^{(\mathbf{w})}_{\mathcal{F}}(S,T)$,

$$\begin{aligned}\Pr\left\{|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(S)}_{\mathbf{w}}\widehat{f}| > \xi\right\} \leq & \Pr\left\{|\overline{\mathrm{E}}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{\mathbf{w}}\widehat{f}| > \xi'\right\} \\ \leq & 2\Pr\left\{|\mathrm{E}'^{(S)}_{\mathbf{w}}\widehat{f} - \mathrm{E}^{(S)}_{\mathbf{w}}\widehat{f}| > \frac{\xi'}{2}\right\}\end{aligned} \tag{62}$$

with $\xi' = \xi - D^{(\mathbf{w})}_{\mathcal{F}}(S,T)$. This completes the proof. ■

By using the resultant deviation inequality and the symmetrization inequality, we can achieve the proof of Theorem 5.1.

A.3 Proof of Theorem 5.1

Proof of Theorem 5.1. Consider ϵ as an independent Rademacher random variables, *i.e.*, an independent $\{-1,1\}$-valued random variable with equal probability of taking either value. Given sample sets $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$, denote for any $f \in \mathcal{F}$ and $1 \leq k \leq K$,

$$\overrightarrow{\epsilon}^{(k)} := (\epsilon_1^{(k)}, \cdots, \epsilon_{N_k}^{(k)}, -\epsilon_1^{(k)}, \cdots, -\epsilon_{N_k}^{(k)}) \in \{-1,1\}^{2N_k}, \tag{63}$$

and for any $f \in \mathcal{F}$,

$$\overrightarrow{f}(\mathbf{Z}_1^{2N_k}) := \big(f(\mathbf{z}'^{(k)}_1), \cdots, f(\mathbf{z}'^{(k)}_{N_k}), f(\mathbf{z}_1^{(k)}), \cdots, f(\mathbf{z}_{N_k}^{(k)})\big). \tag{64}$$

According to (6), (7) and Theorem A.2, given an arbitrary $\xi > D^{(\mathbf{w})}_{\mathcal{F}}(S,T)$, we have for any $\{N_k\}_{k=1}^K \in \mathbb{N}^K$ such that $\prod_{k=1}^K N_k \geq \frac{8(b-a)^2}{(\xi')^2}$ with $\xi' = \xi - D^{(\mathbf{w})}_{\mathcal{F}}(S,T)$,

$$\begin{aligned}&\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}^{(S)}_{\mathbf{w}}f\right| > \xi\right\} \\ \leq & 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}'^{(S)}_{\mathbf{w}}f - \mathrm{E}^{(S)}_{\mathbf{w}}f\right| > \frac{\xi'}{2}\right\} \qquad \text{(by Theorem A.2)} \\ = & 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\frac{w_k}{N_k}\sum_{n=1}^{N_k}\big(f(\mathbf{z}'^{(k)}_n) - f(\mathbf{z}_n^{(k)})\big)\right| > \frac{\xi'}{2}\right\} \\ = & 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\frac{w_k}{N_k}\sum_{n=1}^{N_k}\epsilon_n^{(k)}\big(f(\mathbf{z}'^{(k)}_n) - f(\mathbf{z}_n^{(k)})\big)\right| > \frac{\xi'}{2}\right\} \\ = & 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\sum_{k=1}^{K}\frac{w_k}{2N_k}\big\langle\overrightarrow{\epsilon}^{(k)}, \overrightarrow{f}(\mathbf{Z}_1^{2N_k})\big\rangle\right| > \frac{\xi'}{4}\right\}. \qquad \text{(by (63) and (64))}\end{aligned} \tag{65}$$

Fix a realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$ and let Λ be a $\xi'/8$-radius cover of $\mathcal{F}$ with respect to the $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$ norm. Since $\mathcal{F}$ is composed of the bounded functions with the range $[a,b]$, we assume that the same holds for any $h \in \Lambda$. If f_0 is the function that achieves the following supremum

$$\sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{f}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{4},$$

there must be an $h_0 \in \Lambda$ that satisfies

$$\sum_{k=1}^K \frac{w_k}{2N_k} \Big(|f_0(\mathbf{z'}_n^{(k)}) - h_0(\mathbf{z'}_n^{(k)})| + |f_0(\mathbf{z}_n^{(k)}) - h_0(\mathbf{z}_n^{(k)})| \Big) < \frac{\xi'}{8},$$

and meanwhile,

$$\Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{h_0}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{8}.$$

Therefore, for the realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$, we arrive at

$$\begin{aligned} &\Pr\left\{ \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{f}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{4} \right\} \\ \leq &\Pr\left\{ \sup_{h\in\Lambda} \Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{h}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{8} \right\}. \end{aligned} \tag{66}$$

Moreover, we denote the event

$$A := \left\{ \Pr\left\{ \sup_{h\in\Lambda} \Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{h}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{8} \right\} \right\},$$

and let $\mathbf{1}_A$ be the characteristic function of the event A. By Fubini's Theorem, we have

$$\begin{aligned} \Pr\{A\} &= \mathrm{E}\Big\{ \mathrm{E}_{\overrightarrow{\epsilon}}\big\{\mathbf{1}_A\big\} \Big| \{\mathbf{Z}_1^{2N_k}\}_{k=1}^K \Big\} \\ &= \mathrm{E}\left\{ \Pr\left\{ \sup_{h\in\Lambda} \Big| \sum_{k=1}^K \frac{w_k}{2N_k} \big\langle \overrightarrow{\epsilon}^{(k)}, \overrightarrow{h}(\mathbf{Z}_1^{2N_k}) \big\rangle \Big| > \frac{\xi'}{8} \right\} \Big| \{\mathbf{Z}_1^{2N_k}\}_{k=1}^K \right\}. \end{aligned} \tag{67}$$

Fix a realization of $\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K$ again. According to (63), (64) and Theorem A.1, we have

$$\begin{aligned}
&\Pr\left\{\sup_{h\in\Lambda}\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\big\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\big\rangle\Big|>\frac{\xi'}{8}\right\}\\
\leq&|\Lambda|\max_{h\in\Lambda}\Pr\left\{\Big|\sum_{k=1}^K\frac{w_k}{2N_k}\big\langle\overrightarrow{\epsilon}^{(k)},\overrightarrow{h}(\mathbf{Z}_1^{2N_k})\big\rangle\Big|>\frac{\xi'}{8}\right\}\\
=&\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{\big|\mathrm{E}'^{(S)}_{\mathbf{w}}h-\mathrm{E}^{(S)}_{\mathbf{w}}h\big|>\frac{\xi'}{4}\right\}\\
\leq&\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}'^{(S)}_{\mathbf{w}}h|+|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}^{(S)}_{\mathbf{w}}h|>\frac{\xi'}{4}\right\}\\
\leq&2\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\max_{h\in\Lambda}\Pr\left\{\big|\overline{\mathrm{E}}^{(S)}h-\mathrm{E}^{(S)}_{\mathbf{w}}h\big|>\frac{\xi'}{8}\right\}\\
\leq&4\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\Big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\Big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\},
\end{aligned}\tag{68}$$

where the expectation $\overline{\mathrm{E}}^{(S)}$ is defined in (57).

The combination of (65), (66) and (68) leads to the result: given an arbitrary $\xi>D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and for any $\prod_{k=1}^K N_k\geq\frac{8(b-a)^2}{(\xi')^2}$ with $\xi'=\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$,

$$\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\big|\mathrm{E}^{(T)}f-\mathrm{E}^{(S)}_{\mathbf{w}}f\big|>\xi\right\}\\
\leq&8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\Big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\Big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\}\\
\leq&8\mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F},\xi'/8,2\sum_{k=1}^K N_k\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\Big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\Big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\}.
\end{aligned}\tag{69}$$

According to (69), letting

$$\epsilon:=8\mathcal{N}_1^{\mathbf{w}}\left(\mathcal{F},\xi'/8,2\sum_{k=1}^K N_k\right)\exp\left\{-\frac{\big(\prod_{k=1}^K N_k\big)\Big(\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)\Big)^2}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}\right\},$$

we then arrive at with probability at least $1-\epsilon$,

$$\sup_{f\in\mathcal{F}}\big|\mathrm{E}^{(S)}_{\mathbf{w}}f-\mathrm{E}^{(T)}f\big|\leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T)+\left(\frac{\ln\mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F},\xi'/8,2\sum_{k=1}^K N_k\big)-\ln(\epsilon/8)}{\frac{\big(\prod_{k=1}^K N_k\big)}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k}N_i)\big)}}\right)^{\frac{1}{2}},$$

where $\xi'=\xi-D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. This completes the proof. ■

B Proof of Theorem 6.1

Here, we provide the proof of Theorem 6.1. Similar to the situation of domain adaptation with multiple sources, we need to develop the related Hoeffding-type deviation inequality and the symmetrization inequality for domain adaptation combining source and target data.

B.1 Hoeffding-Type Deviation Inequality

Based on Hoeffding's inequality (Hoeffding, 1963), we derive a deviation inequality for the combination of the source and the target domains.

Theorem B.1 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a, b]$. Let $\mathbf{Z}_1^{N_S} := \{\mathbf{z}_n^{(S)}\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T} := \{\mathbf{z}_n^{(T)}\}_{n=1}^{N_T}$ be sets of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S)} \subset \mathbb{R}^L$ and the target domain $\mathcal{Z}^{(T)} \subset \mathbb{R}^L$, respectively. For any $\tau \in [0, 1)$, define a function $F_\tau : \mathbb{R}^{L(N_S+N_T)} \to \mathbb{R}$ as*

$$F_\tau\left(\mathbf{X}_1^{N_T}, \mathbf{Y}_1^{N_S}\right) := \tau N_S \sum_{n=1}^{N_T} f(\mathbf{x}_n) + (1-\tau) N_T \sum_{n=1}^{N_S} f(\mathbf{y}_n), \tag{70}$$

where

$$\mathbf{X}_1^{N_T} := \{\mathbf{x}_1, \cdots, \mathbf{x}_{N_T}\} \in (\mathbb{R}^L)^{N_T}; \;\; \mathbf{Y}_1^{N_S} := \{\mathbf{y}_1, \cdots, \mathbf{y}_{N_S}\} \in (\mathbb{R}^L)^{N_S}.$$

Then, we have for any $\tau \in [0, 1)$ and any $\xi > 0$,

$$\begin{aligned} &\Pr\left\{\left|F_\tau\left(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\right) - \mathrm{E}^{(*)} F_\tau\right| > \xi\right\} \\ \leq & 2\exp\left\{-\frac{2\xi^2}{(b-a)^2 N_S N_T \left((1-\tau)^2 N_T + \tau^2 N_S\right)}\right\}, \end{aligned} \tag{71}$$

where the expectation $\mathrm{E}^{(*)}$ is taken on both of the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$.

In this theorem, we present a deviation inequality for the combination of source and target domains, which is an extension of the classical Hoeffding-type deviation inequality under the assumption of same distribution (see Bousquet *et al.*, 2004, Theorem 1). Compare to the classical result, the resultant deviation inequality (71) allows the random variables to take values from different domains. The two inequalities coincide when the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ match, *i.e.,* $D_{\mathcal{F}}(S, T) = 0$.

The proof of Theorem B.1 is also processed by a martingale method. Before the formal proof, we introduce some essential notations.

For any $\tau \in [0, 1)$, we denote

$$F_S(\mathbf{Z}_1^{N_S}) := (1-\tau) N_T \sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)}); \;\; F_T(\overline{\mathbf{Z}}_1^{N_T}) := \tau N_S \sum_{n=1}^{N_T} f(\mathbf{z}_n^{(T)}). \tag{72}$$

Recalling (70), it is evident that $F_\tau\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big) = F_S(\mathbf{Z}_1^{N_S}) + F_T(\overline{\mathbf{Z}}_1^{N_T})$. We then define two random variables:

$$\begin{aligned} S_n :=& \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^n\right\},\ 0 \le n \le N_S; \\ T_n :=& \mathrm{E}^{(T)}\left\{F_T(\overline{\mathbf{Z}}_1^{N_T})|\overline{\mathbf{Z}}_1^n\right\},\ 0 \le n \le N_T, \end{aligned} \tag{73}$$

where

$$\begin{aligned} \mathbf{Z}_1^n &= \{\mathbf{z}_1^{(S)}, \cdots, \mathbf{z}_n^{(S)}\} \subseteq \mathbf{Z}_1^{N_S} \ \text{ with } \ \mathbf{Z}_1^0 := \varnothing; \\ \overline{\mathbf{Z}}_1^n &= \{\mathbf{z}_1^{(T)}, \cdots, \mathbf{z}_n^{(T)}\} \subseteq \overline{\mathbf{Z}}_1^{N_T} \ \text{ with } \ \overline{\mathbf{Z}}_1^0 := \varnothing. \end{aligned}$$

It is clear that $S_0 = \mathrm{E}^{(S)}F_S$; $S_{N_S} = F_S(\mathbf{Z}_1^{N_S})$ and $T_0 = \mathrm{E}^{(T)}F_T$; $T_{N_T} = F_T(\overline{\mathbf{Z}}_1^{N_T})$.

According to (70) and (73), we have for any $1 \le n \le N_S$ and any $\tau \in [0, 1)$,

$$\begin{aligned} & S_n - S_{n-1} \\ =& \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^n\right\} - \mathrm{E}^{(S)}\left\{F_S(\mathbf{Z}_1^{N_S})|\mathbf{Z}_1^{n-1}\right\} \\ =& \mathrm{E}^{(S)}\left\{(1-\tau)N_T\sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)})\Big|\mathbf{Z}_1^n\right\} - \mathrm{E}^{(S)}\left\{(1-\tau)N_T\sum_{n=1}^{N_S} f(\mathbf{z}_n^{(S)})\Big|\mathbf{Z}_1^{n-1}\right\} \\ =& (1-\tau)N_T\sum_{m=1}^{n} f(\mathbf{z}_m^{(S)}) + \mathrm{E}^{(S)}\left\{(1-\tau)N_T\sum_{m=n+1}^{N_S} f(\mathbf{z}_m^{(S)})\right\} \\ & - \left((1-\tau)N_T\sum_{m=1}^{n-1} f(\mathbf{z}_m^{(S)}) + \mathrm{E}^{(S)}\left\{(1-\tau)N_T\sum_{m=n}^{N_S} f(\mathbf{z}_m^{(S)})\right\}\right) \\ =& (1-\tau)N_T\left(f(\mathbf{z}_n^{(S)}) - \mathrm{E}^{(S)}f\right). \end{aligned} \tag{74}$$

Similarly, we also have for any $1 \le n \le N_T$,

$$T_n - T_{n-1} = \tau N_S\left(f(\mathbf{z}_n^{(T)}) - \mathrm{E}^{(T)}f\right). \tag{75}$$

We are now ready to prove Theorem B.1.

Proof of Theorem B.1. According to (70) and (72), we have

$$\begin{aligned} F_\tau(\mathbf{Z}_1^N) - \mathrm{E}^{(*)}F_\tau =& F_S(\mathbf{Z}_1^{N_S}) + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(*)}\{F_S + F_T\} \\ =& F_S(\mathbf{Z}_1^{N_S}) - \mathrm{E}^{(S)}F_S + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(T)}F_T. \end{aligned} \tag{76}$$

According to Lemma A.1, (74), (75), (76), Markov's inequality, Jensen's inequality and the law of iterated expectation, we have for any $\alpha > 0$ and any $\tau \in [0, 1)$,

$$
\begin{aligned}
&\Pr\left\{F_\tau(\mathbf{Z}_1^N) - \mathrm{E}^{(*)}F_\tau > \xi\right\} \\
=&\Pr\left\{F_S(\mathbf{Z}_1^{N_S}) - \mathrm{E}^{(S)}F_S + F_T(\overline{\mathbf{Z}}_1^{N_T}) - \mathrm{E}^{(T)}F_T > \xi\right\} \\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(F_S(\mathbf{Z}_1^{N_S})-\mathrm{E}^{(S)}F_S+F_T(\overline{\mathbf{Z}}_1^{N_T})-\mathrm{E}^{(T)}F_T\right)}\right\} \\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S}(S_n-S_{n-1})+\sum_{n=1}^{N_T}(T_n-T_{n-1})\right)}\right\} \\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S}(S_n-S_{n-1})+\sum_{n=1}^{N_T}(T_n-T_{n-1})\right)}\middle|\mathbf{Z}_1^{N_S-1}\right\}\right\} \\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S-1}(S_n-S_{n-1})+\sum_{n=1}^{N_T}(T_n-T_{n-1})\right)}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(S_{N_S}-S_{N_S-1}\right)}\middle|\mathbf{Z}_1^{N_S-1}\right\}\right\} \\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S-1}(S_n-S_{n-1})+\sum_{n=1}^{N_T}(T_n-T_{n-1})\right)}\right\}\mathrm{e}^{\frac{(1-\tau)^2N_T^2\alpha^2(b-a)^2}{8}} \\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S-1}(S_n-S_{n-1})+\sum_{n=1}^{N_T-1}(T_n-T_{n-1})\right)}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(T_{N_T}-T_{N_T-1}\right)}\middle|\overline{\mathbf{Z}}_1^{N_T-1}\right\}\right\} \\
&\times\mathrm{e}^{\frac{(1-\tau)^2N_T^2\alpha^2(b-a)^2}{8}} \\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{n=1}^{N_S-1}(S_n-S_{n-1})+\sum_{n=1}^{N_T-1}(T_n-T_{n-1})\right)}\right\}\mathrm{e}^{\frac{\tau^2N_S^2\alpha^2(b-a)^2}{8}}\mathrm{e}^{\frac{(1-\tau)^2N_T^2\alpha^2(b-a)^2}{8}}. \qquad (77)
\end{aligned}
$$

Then, we have

$$\Pr\left\{F_\tau\left(\mathbf{Z}_1^{N_S},\overline{\mathbf{Z}}_1^{N_T}\right) - \mathrm{E}^{(*)}F_\tau > \xi\right\} \leq \mathrm{e}^{\Phi(\alpha)-\alpha\xi}, \qquad (78)$$

where

$$\Phi(\alpha) = \frac{\alpha^2(1-\tau)^2(b-a)^2N_SN_T^2}{8} + \frac{\alpha^2\tau^2(b-a)^2N_S^2N_T}{8}. \qquad (79)$$

Similarly, we can arrive at

$$\Pr\left\{\mathrm{E}^{(*)}F_\tau - F_\tau\left(\mathbf{Z}_1^{N_S},\overline{\mathbf{Z}}_1^{N_T}\right) > \xi\right\} \leq \mathrm{e}^{\Phi(\alpha)-\alpha\xi}. \qquad (80)$$

Note that $\Phi(\alpha) - \alpha\xi$ is a quadratic function with respect to $\alpha > 0$ and thus the minimum value

$$\min_{\alpha>0}\left\{\Phi(\alpha) - \alpha\xi\right\}$$

is achieved when

$$\alpha = \frac{4\xi}{(b-a)^2N_SN_T\left((1-\tau)^2N_T + \tau^2N_S\right)}.$$

By combining (78), (79) and (80), we arrive at

$$\Pr\left\{\left|F_\tau\left(\mathbf{Z}_1^{N_S},\overline{\mathbf{Z}}_1^{N_T}\right) - \mathrm{E}^{(*)}F_\tau\right| > \xi\right\} \leq 2\exp\left\{-\frac{2\xi^2}{(b-a)^2N_SN_T\left((1-\tau)^2N_T + \tau^2N_S\right)}\right\}.$$

This completes the proof. ■

B.2 Symmetrization Inequality

In the following theorem, we present the symmetrization inequality for domain adaptation combining source and target data.

Theorem B.2 *Assume that $\mathcal{F}$ is a function class with the range $[a,b]$. Let $\mathbf{Z}_1^{N_S}$ and $\mathbf{Z'}_1^{N_S}$ be drawn from the source domain $\mathcal{Z}^{(S)}$, and $\overline{\mathbf{Z}}_1^{N_T}$ and $\overline{\mathbf{Z}'}_1^{N_T}$ be drawn from the target domain $\mathcal{Z}^{(T)}$. Then, for any $\tau \in [0,1)$ and given an arbitrary $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$,*

$$\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right| > \xi\right\} \leq 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E'}_\tau f - \mathrm{E}_\tau f\right| > \frac{\xi'}{2}\right\} \tag{81}$$

with $\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T)$.

This theorem shows that for any $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$, the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right| > \xi$$

can be bounded by using the probability of the event:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E'}_\tau f - \mathrm{E}_\tau f\right| > \frac{\xi'}{2}$$

that is only determined by the samples drawn from the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$, when $N_S N_T \geq \frac{8(b-a)^2}{(\xi')^2}$. Compared to the classical symmetrization result under the assumption of same distribution (see Bousquet *et al.*, 2004), there is a discrepancy term $(1-\tau)D_{\mathcal{F}}(S,T)$. The two results will coincide when the source and the target domains match, *i.e.*, $D_{\mathcal{F}}(S,T) = 0$. The following is the proof of Theorem B.2.

Proof of Theorem B.2. Let $\widehat{f}$ be the function achieving the supremum:

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right|$$

with respect to $\mathbf{Z}_1^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T}$. According to (9) and (12), we arrive at

$$\begin{aligned}
\left|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_\tau\widehat{f}\right| =& \left|\tau\mathrm{E}^{(T)}\widehat{f} + (1-\tau)\mathrm{E}^{(T)}\widehat{f} - (1-\tau)\mathrm{E}^{(S)}\widehat{f} + (1-\tau)\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}_\tau\widehat{f}\right| \\
=& \left|\tau(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(S)}\widehat{f}) + (1-\tau)(\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| \\
\leq& (1-\tau)D_{\mathcal{F}}(S,T) + \left|\tau(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right|,
\end{aligned} \tag{82}$$

and thus

$$\begin{aligned}
&\Pr\left\{\left|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_\tau\widehat{f}\right| > \xi\right\} \\
\leq&\Pr\left\{(1-\tau)D_{\mathcal{F}}(S,T) + \left|\tau(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| > \xi\right\},
\end{aligned} \tag{83}$$

where

$$\mathrm{E}_{N_T}^{(T)}\widehat{f} := \frac{1}{N_T}\sum_{n=1}^{N_T}\widehat{f}(\mathbf{z}_n^{(T)}); \quad \mathrm{E}_{N_S}^{(S)}\widehat{f} := \frac{1}{N_S}\sum_{n=1}^{N_S}\widehat{f}(\mathbf{z}_n^{(S)}). \tag{84}$$

Let

$$\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T) \tag{85}$$

and denote $\wedge$ as the conjunction of two events. According to the triangle inequality, we have

$$\begin{aligned}
&\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}_{N_T}^{(T)}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}_{N_S}^{(S)}\widehat{f})|>\xi'\right\}}\right)\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})|<\frac{\xi'}{2}\right\}}\right)\\
=&\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}_{N_T}^{(T)}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}_{N_S}^{(S)}\widehat{f})|>\xi'\right\}\wedge\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f})-\mathrm{E}'^{(S)}_{N_S}\widehat{f}|<\frac{\xi'}{2}\right\}}\\
\leq&\mathbf{1}_{\left\{|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}_{N_T}^{(T)}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}_{N_S}^{(S)}\widehat{f})|>\frac{\xi'}{2}\right\}}.
\end{aligned}$$

Then, taking the expectation with respect to ${\mathbf{Z}'}_1^{N_S}$ and $\overline{\mathbf{Z}}'^{N_T}_1$ gives

$$\begin{aligned}
&\left(\mathbf{1}_{\left\{|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}_{N_T}^{(T)}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}_{N_S}^{(S)}\widehat{f})|>\xi'\right\}}\right)\\
&\times\Pr'\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})\right|<\frac{\xi'}{2}\right\}\\
&\leq\Pr'\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f}-\mathrm{E}_{N_T}^{(T)}\widehat{f})+(1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f}-\mathrm{E}_{N_S}^{(S)}\widehat{f})\right|>\frac{\xi'}{2}\right\}.
\end{aligned}\tag{86}$$

By Chebyshev's inequality, since ${\mathbf{Z}'}_1^{N_S}=\{\mathbf{z}'^{(S)}_n\}_{n=1}^{N_S}$ and $\overline{\mathbf{Z}}'^{N_T}_1=\{\mathbf{z}'^{(T)}_n\}_{n=1}^{N_T}$ are sets of i.i.d. samples drawn from the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$ respectively, we have for any $\xi'>0$ and any $\tau\in[0,1)$,

$$\begin{aligned}
&\Pr'\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f}-\mathrm{E}'^{(T)}_{N_T}\widehat{f})+(1-\tau)(\mathrm{E}^{(S)}\widehat{f}-\mathrm{E}'^{(S)}_{N_S}\widehat{f})\right|\geq\frac{\xi'}{2}\right\}\\
\leq&\Pr'\left\{\frac{\tau}{N_T}\sum_{n=1}^{N_T}|\mathrm{E}^{(T)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(T)}_n)|+\frac{1-\tau}{N_S}\sum_{n=1}^{N_S}|\mathrm{E}^{(S)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(S)}_n)|\geq\frac{\xi'}{2}\right\}\\
\leq&\frac{4\mathrm{E}\left\{\tau N_S N_T(\mathrm{E}^{(T)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(T)}))^2+(1-\tau)N_S N_T(\mathrm{E}^{(S)}\widehat{f}-\widehat{f}(\mathbf{z}'^{(S)}))^2\right\}}{N_S^2N_T^2(\xi')^2}\\
\leq&\frac{4\mathrm{E}\left\{\tau N_S N_T(b-a)^2+(1-\tau)N_S N_T(b-a)^2\right\}}{N_S^2N_T^2(\xi')^2}\\
=&\frac{4(b-a)^2}{N_S N_T(\xi')^2},
\end{aligned}\tag{87}$$

where $\mathbf{z}'^{(S)}$ and $\mathbf{z}'^{(T)}$ stand for the ghost random variables taking values from the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$, respectively.

Subsequently, according to (86) and (87), we have for any $\xi' > 0$,

$$\begin{aligned}
&\Pr'\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| > \frac{\xi'}{2}\right\} \\
&\geq \left(\mathbf{1}_{\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| > \xi'\right\}}\right)\left(1 - \frac{4(b-a)^2}{N_S N_T(\xi')^2}\right). \qquad (88)
\end{aligned}$$

According to (83), (85) and (88), by letting

$$\frac{4(b-a)^2}{N_S N_T(\xi')^2} \leq \frac{1}{2},$$

and taking the expectation with respect to $\mathbf{Z}_1^{N_S}$ and $\overline{\mathbf{Z}}_1^{N_T}$, we have for any $\xi' > 0$,

$$\begin{aligned}
&\Pr\left\{\left|\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}_\tau\widehat{f}\right| > \xi\right\} \\
\leq &\Pr\left\{\left|\tau(\mathrm{E}^{(T)}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}^{(S)}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| > \xi'\right\} \\
\leq &2\Pr\left\{\left|\tau(\mathrm{E}'^{(T)}_{N_T}\widehat{f} - \mathrm{E}^{(T)}_{N_T}\widehat{f}) + (1-\tau)(\mathrm{E}'^{(S)}_{N_S}\widehat{f} - \mathrm{E}^{(S)}_{N_S}\widehat{f})\right| > \frac{\xi'}{2}\right\} \\
= &2\Pr\left\{\left|\mathrm{E}'_\tau\widehat{f} - \mathrm{E}_\tau\widehat{f}\right| > \frac{\xi'}{2}\right\} \qquad (89)
\end{aligned}$$

with $\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T)$. This completes the proof. ■

We are now ready to prove Theorem 6.1.

B.3 Proof of Theorem 6.1

Proof of Theorem 6.1. Consider $\{\epsilon_n\}_{n=1}^N$ as independent Rademacher random variables, *i.e.*, independent $\{\pm 1\}$-valued random variables with equal probability of taking either value. Given $\{\epsilon_n\}_{n=1}^{N_S}$, $\{\epsilon_n\}_{n=1}^{N_T}$, $\mathbf{Z}_1^{2N_S}$ and $\overline{\mathbf{Z}}_1^{2N_T}$, denote

$$\begin{aligned}
\overrightarrow{\epsilon}_S :=&(\epsilon_1, \cdots, \epsilon_{N_S}, -\epsilon_1, \cdots, -\epsilon_{N_S}) \in \{\pm 1\}^{2N_S}; \\
\overrightarrow{\epsilon}_T :=&(\epsilon_1, \cdots, \epsilon_{N_T}, -\epsilon_1, \cdots, -\epsilon_{N_T}) \in \{\pm 1\}^{2N_T}, \qquad (90)
\end{aligned}$$

and for any $f \in \mathcal{F}$,

$$\begin{aligned}
\overrightarrow{f}(\mathbf{Z}_1^{2N_S}) :=&\big(f(\mathbf{z}'_1), \cdots, f(\mathbf{z}'_{N_S}), f(\mathbf{z}_1), \cdots, f(\mathbf{z}_{N_S})\big) \in [a,b]^{2N_S}; \\
\overrightarrow{f}(\mathbf{Z}_1^{2N_T}) :=&\big(f(\mathbf{z}'_1), \cdots, f(\mathbf{z}'_{N_T}), f(\mathbf{z}_1), \cdots, f(\mathbf{z}_{N_T})\big) \in [a,b]^{2N_T}. \qquad (91)
\end{aligned}$$

We also denote

$$\begin{aligned}
\mathbf{Z} :=&\left\{\overline{\mathbf{Z}}_1^{2N_T}, \mathbf{Z}_1^{2N_S}\right\} \in \left(\mathcal{Z}^{(T)}\right)^{2N_T} \times \left(\mathcal{Z}^{(S)}\right)^{2N_S}; \\
\overrightarrow{\epsilon} :=&\Big(\underbrace{\overrightarrow{\epsilon}_T, \cdots, \overrightarrow{\epsilon}_T}_{N_S}, \underbrace{\overrightarrow{\epsilon}_S, \cdots, \overrightarrow{\epsilon}_S}_{N_T}\Big) \in \{\pm 1\}^{4N_S N_T}; \\
\overrightarrow{f}(\mathbf{Z}) :=&\Big(\underbrace{\overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T}), \cdots, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})}_{N_S}, \underbrace{\overrightarrow{f}(\mathbf{Z}_1^{2N_S}), \cdots, \overrightarrow{f}(\mathbf{Z}_1^{2N_S})}_{N_T}\Big) \in [a,b]^{4N_S N_T}. \qquad (92)
\end{aligned}$$

According to (6), (85) and Theorem B.2, for any $\tau \in [0,1)$ and given an arbitrary $\xi > (1-\tau)D_{\mathcal{F}}(S,T)$, we have for any $N_S N_T \geq \frac{8(b-a)^2}{\xi'^2}$ with $\xi' = \xi - (1-\tau)D_{\mathcal{F}}(S,T)$,

$$
\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f - \mathrm{E}_\tau f\right| > \xi\right\} \\
\leq& 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}'_\tau f - \mathrm{E}_\tau f\right| > \frac{\xi'}{2}\right\} \qquad \text{(by Theorem B.2)} \\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\frac{\tau}{N_T}\sum_{n=1}^{N_T}\left(f(\mathbf{z'}_n^{(T)}) - f(\mathbf{z}_n^{(T)})\right) + \frac{1-\tau}{N_S}\sum_{n=1}^{N_S}\left(f(\mathbf{z'}_n^{(S)}) - f(\mathbf{z}_n^{(S)})\right)\right| > \frac{\xi'}{2}\right\} \\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\frac{\tau}{N_T}\sum_{n=1}^{N_T}\epsilon_n\left(f(\mathbf{z'}_n^{(T)}) - f(\mathbf{z}_n^{(T)})\right) + \frac{1-\tau}{N_S}\sum_{n=1}^{N_S}\epsilon_n\left(f(\mathbf{z'}_n^{(S)}) - f(\mathbf{z}_n^{(S)})\right)\right| > \frac{\xi'}{2}\right\} \\
=& 2\Pr\left\{\sup_{f\in\mathcal{F}}\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{f}(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{4}\right\}. \qquad (93)
\end{aligned}
$$

Given a $\tau \in [0,1)$, fix a realization of $\mathbf{Z}$ and let Λ be a $\xi'/8$-radius cover of $\mathcal{F}$ with respect to the $\ell_1^\tau(\mathbf{Z})$ norm. Since $\mathcal{F}$ is composed of bounded functions with the range $[a,b]$, we assume that the same holds for any $h \in \Lambda$. If f_0 is the function that achieves the following supremum

$$
\sup_{f\in\mathcal{F}}\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{f}(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{4},
$$

there must be an $h_0 \in \Lambda$ that satisfies that

$$
\begin{aligned}
&\frac{\tau}{2N_T}\sum_{n=1}^{N_T}\left(|f_0(\mathbf{z'}_n^{(T)}) - h_0(\mathbf{z'}_n^{T})| + |f_0(\mathbf{z}_n^{(T)}) - h_0(\mathbf{z}_n^{(T)})|\right) \\
&+ \frac{1-\tau}{2N_S}\sum_{n=1}^{N_S}\left(|f_0(\mathbf{z'}_n^{(S)}) - h_0(\mathbf{z'}_n^{S})| + |f_0(\mathbf{z}_n^{(S)}) - h_0(\mathbf{z}_n^{(S)})|\right) < \frac{\xi'}{8},
\end{aligned}
$$

and meanwhile,

$$
\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{h}_0(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{h}_0(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{8}.
$$

Therefore, for the realization of $\mathbf{Z}$, we arrive at

$$
\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{f}(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{f}(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{4}\right\} \\
\leq& \Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{h}(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{8}\right\}. \qquad (94)
\end{aligned}
$$

Moreover, we denote the event

$$
A := \left\{\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\left\langle\overrightarrow{\epsilon}_T, \overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\right\rangle + \frac{1-\tau}{2N_S}\left\langle\overrightarrow{\epsilon}_S, \overrightarrow{h}(\mathbf{Z}_1^{2N_S})\right\rangle\right| > \frac{\xi'}{8}\right\}\right\},
$$

and let $\mathbf{1}_A$ be the characteristic function of the event A. By Fubini's Theorem, we have

$$\begin{aligned}
&\Pr\{A\} = \mathrm{E}\Big\{\mathrm{E}_{\overrightarrow{\epsilon}}\big\{\mathbf{1}_A\big\}\big|\ \mathbf{Z}\Big\}\\
=&\mathrm{E}\left\{\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\big\langle\overrightarrow{\epsilon}_T,\,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\big\rangle+\frac{1-\tau}{2N_S}\big\langle\overrightarrow{\epsilon}_S,\,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\big\rangle\right|>\frac{\xi'}{8}\right\}\ \middle|\ \mathbf{Z}\right\}. \qquad (95)
\end{aligned}$$

Fix a realization of $\mathbf{Z}$ again. According to (21), (90), (91) and Theorem B.1, for any $\tau\in[0,1)$ and given an arbitrary $\xi'>0$, we have for any $N_SN_T\geq\frac{8(b-a)^2}{(\xi')^2}$,

$$\begin{aligned}
&\Pr\left\{\sup_{h\in\Lambda}\left|\frac{\tau}{2N_T}\big\langle\overrightarrow{\epsilon}_T,\,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\big\rangle+\frac{1-\tau}{2N_S}\big\langle\overrightarrow{\epsilon}_S,\,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\big\rangle\right|>\frac{\xi'}{8}\right\}\\
\leq&|\Lambda|\max_{h\in\Lambda}\Pr\left\{\left|\frac{\tau}{2N_T}\big\langle\overrightarrow{\epsilon}_T,\,\overrightarrow{h}(\overline{\mathbf{Z}}_1^{2N_T})\big\rangle+\frac{1-\tau}{2N_S}\big\langle\overrightarrow{\epsilon}_S,\,\overrightarrow{h}(\mathbf{Z}_1^{2N_S})\big\rangle\right|>\frac{\xi'}{8}\right\}\\
=&\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\mathrm{E'}_\tau h-\mathrm{E}_\tau h\right|>\frac{\xi'}{4}\right\}\\
\leq&\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\widetilde{\mathrm{E}}h-\mathrm{E'}_\tau h\right|+\left|\widetilde{\mathrm{E}}h-\mathrm{E}_\tau h\right|>\frac{\xi'}{4}\right\}\\
\leq&2\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\max_{h\in\Lambda}\Pr\left\{\left|\widetilde{\mathrm{E}}h-\mathrm{E}_\tau h\right|>\frac{\xi'}{8}\right\}\\
\leq&4\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\exp\left\{-\frac{N_SN_T\left(\xi-(1-\tau)D_{\mathcal{F}}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2N_T+\tau^2N_S\right)}\right\}, \qquad (96)
\end{aligned}$$

where $\widetilde{\mathrm{E}}h:=\tau\mathrm{E}^{(T)}h+(1-\tau)\mathrm{E}^{(S)}h$.

The combination of (22), (93), (94) and (96) leads to the following result: for any $\tau\in[0,1)$ and given an arbitrary $\xi>(1-\tau)D_{\mathcal{F}}(S,T)$, we have for any $N_SN_T\geq\frac{8(b-a)^2}{(\xi')^2}$,

$$\begin{aligned}
&\Pr\left\{\sup_{f\in\mathcal{F}}\left|\mathrm{E}^{(T)}f-\mathrm{E}_\tau f\right|>\xi\right\}\\
\leq&8\mathrm{E}\mathcal{N}\left(\mathcal{F},\xi'/8,\ell_1^\tau(\mathbf{Z})\right)\exp\left\{-\frac{N_SN_T\left(\xi-(1-\tau)D_{\mathcal{F}}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2N_T+\tau^2N_S\right)}\right\}\\
\leq&8\mathcal{N}_1^\tau(\mathcal{F},\xi'/8,2(N_S+N_T))\exp\left\{-\frac{N_SN_T\left(\xi-(1-\tau)D_{\mathcal{F}}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2N_T+\tau^2N_S\right)}\right\}. \qquad (97)
\end{aligned}$$

According to (97), letting

$$\epsilon:=8\mathcal{N}_1^\tau(\mathcal{F},\xi'/8,2(N_S+N_T))\exp\left\{-\frac{N_SN_T\left(\xi-(1-\tau)D_{\mathcal{F}}(S,T)\right)^2}{32(b-a)^2\left((1-\tau)^2N_T+\tau^2N_S\right)}\right\},$$

we have given an arbitrary $\xi>(1-\tau)D_{\mathcal{F}}(S,T)$ and for any $N_SN_T\geq\frac{8(b-a)^2}{(\xi')^2}$, with probability at least $1-\epsilon$,

$$\sup_{f\in\mathcal{F}}\left|\mathrm{E}_\tau f-\mathrm{E}^{(T)}f\right|\leq(1-\tau)D_{\mathcal{F}}(S,T)+\left(\frac{\ln\mathcal{N}_1^\tau(\mathcal{F},\xi'/8,2(N_S+N_T))-\ln(\epsilon/8)}{\frac{N_SN_T}{32(b-a)^2((1-\tau)^2N_T+\tau^2N_S)}}\right)^{\frac{1}{2}}.$$

This completes the proof. ■

C Proofs of Theorems 5.2 & 6.2

In this appendix, we will prove Theorem 5.2 and Theorem 6.2. In order to achieve the proofs, we need to generalize the classical McDiarmid's inequality (see Bousquet *et al.*, 2004, Theorem 6) to a more general setting where independent random variables can independently take values from different domains.

C.1 Generalized McDiarmid's inequality

The following is the classical McDiarmid's inequality that is one of the most frequently used deviation inequalities in statistical learning theory and has been widely used to obtain generalization bounds based on the Rademacher complexity under the assumption of same distribution (see Bousquet *et al.*, 2004, Theorem 6).

Theorem C.1 (McDiamid's Inequality) *Let $\mathbf{z}_1, \cdots, \mathbf{z}_N$ be N independent random variables taking value from the domain $\mathcal{Z}$. Assume that the function $H : \mathcal{Z} \to \mathbb{R}$ satisfies the condition of bounded difference: for all $1 \leq n \leq N$,*

$$\sup_{\mathbf{z}_1, \cdots, \mathbf{z}_N, \mathbf{z'}_n} \Big| H\big(\mathbf{z}_1, \cdots, \mathbf{z}_n, \cdots, \mathbf{z}_N\big) - H\big(\mathbf{z}_1, \cdots, \mathbf{z'}_n, \cdots, \mathbf{z}_N\big) \Big| \leq c_n. \tag{98}$$

Then, for any $\xi > 0$

$$\Pr\Big\{ H\big(\mathbf{z}_1, \cdots, \mathbf{z}_n, \cdots, \mathbf{z}_N\big) - \mathrm{E}\Big\{ H\big(\mathbf{z}_1, \cdots, \mathbf{z}_n, \cdots, \mathbf{z}_N\big) \Big\} \geq \xi \Big\} \leq \exp\left\{ -2\xi^2 / \sum_{n=1}^{N} c_n^2 \right\}.$$

As shown in Theorem C.1, the classical McDiarmid's inequality is valid under the condition that random variables $\mathbf{z}_1, \cdots, \mathbf{z}_N$ are independent and drawn from the same domain. Next, we generalize this inequality to a more general setting, where the independent random variables can take values from different domains.

Theorem C.2 *Given independent domains $\mathcal{Z}^{(S_k)}$ $(1 \leq k \leq K)$, for any $1 \leq k \leq K$, let $\mathbf{Z}_1^{N_k} := \{\mathbf{z}_n^{(S_k)}\}_{n=1}^{N_k}$ be N_k independent random variables taking values from the domain $\mathcal{Z}^{(S_k)}$. Assume that the function $H : \big(\mathcal{Z}^{(S_1)}\big)^{N_1} \times \cdots \times \big(\mathcal{Z}^{(S_K)}\big)^{N_K} \to \mathbb{R}$ satisfies the condition of bounded difference: for all $1 \leq k \leq K$ and $1 \leq n \leq N_k$,*

$$\begin{aligned} \sup_{\mathbf{z}_1^{N_1}, \cdots, \mathbf{z}_1^{N_K}, \mathbf{z'}_n^{(S_k)}} \Big| & H\big(\mathbf{Z}_1^{N_1}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{z}_1^{(S_k)}, \cdots, \mathbf{z}_n^{(S_k)}, \cdots, \mathbf{z}_{N_k}^{(S_k)}, \mathbf{Z}_1^{N_{k+1}}, \cdots, \mathbf{Z}_1^{N_K}\big) \\ - & H\big(\mathbf{Z}_1^{N_1}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{z}_1^{(S_k)}, \cdots, \mathbf{z'}_n^{(S_k)}, \cdots, \mathbf{z}_{N_k}^{(S_k)}, \mathbf{Z}_1^{N_{k+1}}, \cdots, \mathbf{Z}_1^{N_K}\big) \Big| \leq c_n^{(k)}. \end{aligned} \tag{99}$$

Then, for any $\xi > 0$

$$\Pr\Big\{ H\big(\mathbf{Z}_1^{N_1}, \cdots, \mathbf{Z}_1^{N_K}\big) - \mathrm{E}\Big\{ H\big(\mathbf{Z}_1^{N_1}, \cdots, \mathbf{Z}_1^{N_K}\big) \Big\} \geq \xi \Big\} \leq \exp\left\{ -2\xi^2 / \sum_{k=1}^{K} \sum_{n=1}^{N_k} (c_n^{(k)})^2 \right\}.$$

Proof. Define a random variable

$$T_n^{(k)} := \mathrm{E}\left\{H(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^n\right\}, \quad 1 \le k \le K,\ 0 \le n \le N_k, \tag{100}$$

where

$$\mathbf{Z}_1^n = \{\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \cdots, \mathbf{z}_n^{(k)}\} \subseteq \mathbf{Z}_1^{N_k}, \text{ and } \mathbf{Z}_1^0 = \varnothing.$$

It is clear that

$$T_0^{(1)} = \mathrm{E}\big\{H(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)\big\} \text{ and } T_{N_K}^{(K)} = H(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K),$$

and thus

$$H(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K) - \mathrm{E}\big\{H(\{\mathbf{Z}_1^{N_k}\}_{k=1}^K)\big\} = T_{N_K}^{(K)} - T_0^{(1)} = \sum_{k=1}^K \sum_{n=1}^{N_k} (T_n^{(k)} - T_{n-1}^{(k)}). \tag{101}$$

Denote for any $1 \le k \le K$ and $1 \le n \le N_k$,

$$\begin{aligned} U_n^{(k)} &= \sup_{\mu} \left\{T_n^{(k)}\big|_{\mathbf{z}_n^{(k)}=\mu} - T_{n-1}^{(k)}\right\}; \\ L_n^{(k)} &= \inf_{\nu} \left\{T_n^{(k)}\big|_{\mathbf{z}_n^{(k)}=\nu} - T_{n-1}^{(k)}\right\}. \end{aligned}$$

It follows from the definition of (100) that $L_n^{(k)} \le (T_n^{(k)} - T_{n-1}^{(k)}) \le U_n^{(k)}$ and thus results in

$$T_n^{(k)} - T_{n-1}^{(k)} \le U_n^{(k)} - L_n^{(k)} = \sup_{\mu,\nu} \left\{T_n^{(k)}\big|_{\mathbf{z}_n^{(k)}=\mu} - T_n^{(k)}\big|_{\mathbf{z}_n^{(k)}=\nu}\right\} \le c_n^{(k)}. \tag{102}$$

Moreover, by the law of iterated expectation, we also have for any $1 \le k \le K$ and $1 \le n \le N_k$

$$\mathrm{E}\left\{T_n^{(k)} - T_{n-1}^{(k)}|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1}\right\} = 0. \tag{103}$$

According to Hoeffding inequality (see Hoeffding, 1963), given an $\alpha > 0$, the condition (99) leads to for any $1 \le k \le K$ and $1 \le n \le N_k$,

$$\mathrm{E}\left\{\mathrm{e}^{\alpha(T_n^{(k)} - T_{n-1}^{(k)})}|\mathbf{Z}_1^{N_1}, \mathbf{Z}_1^{N_2}, \cdots, \mathbf{Z}_1^{N_{k-1}}, \mathbf{Z}_1^{n-1}\right\} \le \mathrm{e}^{\alpha^2 (c_n^{(k)})^2/8}. \tag{104}$$

Subsequently, according to Markov's inequality, (101), (102), (103) and (104), we have for any $\alpha > 0$,

$$
\begin{aligned}
&\Pr\left\{H\big(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\big)-\mathrm{E}\Big\{H\big(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\big)\Big\}\geq\xi\right\}\\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\Big\{\mathrm{e}^{\alpha\left(H\left(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\right)-\mathrm{E}\left\{H\left(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\right)\right\}\right)}\Big\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\Big\{\mathrm{e}^{\alpha\left(\sum_{k=1}^{K}\sum_{n=1}^{N_k}(T_n^{(k)}-T_{n-1}^{(k)})\right)}\Big\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\left\{\mathrm{E}\left\{\mathrm{e}^{\alpha\left(\sum_{k=1}^{K}\sum_{n=1}^{N_k}(T_n^{(k)}-T_{n-1}^{(k)})\right)}\right\}|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{k-1}},\mathbf{Z}_1^{N_K-1}\right\}\\
=&\mathrm{e}^{-\alpha\xi}\mathrm{E}\Big\{\mathrm{e}^{\alpha\left(\sum_{k=1}^{K-1}\sum_{n=1}^{N_k}(T_n^{(k)}-T_{n-1}^{(k)})+\sum_{n=1}^{N_K-1}(T_n^{(K)}-T_{n-1}^{(K)})\right)}\Big\{\mathrm{e}^{\alpha\left(T_{N_K}^{(K)}-T_{N_K-1}^{(K)}\right)}|\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_{K-1}},\mathbf{Z}_1^{N_K-1}\Big\}\Big\}\\
\leq&\mathrm{e}^{-\alpha\xi}\mathrm{E}\Big\{\mathrm{e}^{\alpha\left(\sum_{k=1}^{K-1}\sum_{n=1}^{N_k}(T_n^{(k)}-T_{n-1}^{(k)})+\sum_{n=1}^{N_K-1}(T_n^{(K)}-T_{n-1}^{(K)})\right)}\Big\}\mathrm{e}^{\alpha^2(c_{N_K}^{(K)})^2/8}\\
\leq&\mathrm{e}^{-\alpha\xi}\prod_{k=1}^{K}\prod_{n=1}^{N_k}\exp\left\{\frac{\alpha^2(c_n^{(k)})^2}{8}\right\}\\
=&\exp\left\{-\alpha\xi+\alpha^2\sum_{k=1}^{K}\sum_{n=1}^{N_k}\frac{(c_n^{(k)})^2}{8}\right\}.
\end{aligned}
$$

The above bound is minimized by setting

$$\alpha^*=\frac{4\xi}{\sum_{k=1}^{K}\sum_{n=1}^{N_K}(c_n^{(k)})^2},$$

and its minimum value is

$$\exp\left\{-2\xi^2/\sum_{k=1}^{K}\sum_{n=1}^{N_k}(c_n^{(k)})^2\right\}.$$

This completes the proof. ■

C.2 Proof of Theorem 5.2

By using Theorem C.2, we prove Theorem 5.2 as follows:

Proof of Theorem 5.2 Assume that $\mathcal{F}$ is a function class $\mathcal{F}$ consisting of bounded functions with the range $[a,b]$. Let sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K:=\{\{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}\}_{k=1}^K$ be drawn from multiple sources $\mathcal{Z}^{(S_k)}$ $(1\leq k\leq K)$, respectively. Given a choice of $\mathbf{w}\in[0,1]^K$ with $\sum_{k=1}^K w_k=1$, denote

$$H\big(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\big):=\sup_{f\in\mathcal{F}}\big|\mathrm{E}_{\mathbf{w}}^{(S)}f-\mathrm{E}^{(T)}f\big|. \tag{105}$$

By (1), we have

$$H\big(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\big)=\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{K}w_k\big(\mathrm{E}_{N_k}^{(S)}f-\mathrm{E}^{(T)}f\big)\Big|, \tag{106}$$

where $\mathrm{E}_{N_k}^{(S)}f=\frac{1}{N_k}\sum_{n=1}^{N_k}f(\mathbf{z}_n^{(k)})$. Therefore, it is clear that such $H\big(\mathbf{Z}_1^{N_1},\cdots,\mathbf{Z}_1^{N_K}\big)$ satisfies the condition of bounded difference with

$$c_n^{(k)}=\frac{(b-a)w_k}{N_k}$$

for all $1 \le k \le K$ and $1 \le n \le N_k$. Thus, according to Theorem C.2, we have for any $\xi > 0$,

$$\Pr\Big\{H\big(\mathbf{Z}_1^{N_k}, \cdots, \mathbf{Z}_1^{N_k}\big) - \mathrm{E}^{(S)}\big\{H\big(\mathbf{Z}_1^{N_k}, \cdots, \mathbf{Z}_1^{N_k}\big)\big\} \ge \xi\Big\} \le \exp\left\{\frac{-2\xi^2}{\sum_{k=1}^K \frac{(b-a)^2 w_k^2}{N_k}}\right\},$$

which can be equivalently rewritten as with probability at least $1-\epsilon$,

$$\begin{aligned}
H\big(\mathbf{Z}_1^{N_k}, \cdots, \mathbf{Z}_1^{N_k}\big) \le & \mathrm{E}^{(S)}\big\{H\big(\mathbf{Z}_1^{N_k}, \cdots, \mathbf{Z}_1^{N_k}\big)\big\} + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}} \\
= & \mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^K w_k\big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}^{(T)} f\big)\Big|\right\} + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}} \\
= & \mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^K w_k\big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}^{(S_k)} f + \mathrm{E}^{(S_k)} f - \mathrm{E}^{(T)} f\big)\Big|\right\} \\
& + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}} \\
\le & \mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^K w_k\big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}^{(S_k)} f\big)\Big|\right\} + \sum_{k=1}^K w_k \sup_{f\in\mathcal{F}}\big|\mathrm{E}^{(S_k)} f - \mathrm{E}^{(T)} f\big| \\
& + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}} \\
\le & \mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^K w_k\big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}^{(S_k)} f\big)\Big|\right\} + D_{\mathcal{F}}^{(\mathbf{w})}(S,T) \quad \text{[see (26)]} \\
& + \sqrt{\sum_{k=1}^K \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}}. \qquad (107)
\end{aligned}$$

Next, according to (23) and (106), we have

$$
\begin{aligned}
&\mathrm{E}^{(S)} \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}^{(S_k)} f \big) \Big| \\
=&\mathrm{E}^{(S)} \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}'^{(S_k)} \{ \mathrm{E}'^{(S)}_{N_k} f \} \big) \Big| \\
\leq&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \big(\mathrm{E}_{N_k}^{(S)} f - \mathrm{E}'^{(S)}_{N_k} f \big) \Big| \\
=&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \frac{1}{N_k} \sum_{n=1}^{N_k} \big(f(\mathbf{z}_n^{(k)}) - f(\mathbf{z}'^{(k)}_n) \big) \Big| \\
\leq&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \mathrm{E}_\sigma \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \frac{1}{N_k} \sum_{n=1}^{N_k} \sigma_n^{(k)} \big(f(\mathbf{z}_n^{(k)}) - f(\mathbf{z}'^{(k)}_n) \big) \Big| \\
\leq&2\mathrm{E}^{(S)} \mathrm{E}_\sigma \sup_{f\in\mathcal{F}} \Big| \sum_{k=1}^{K} w_k \frac{1}{N_k} \sum_{n=1}^{N_k} \sigma_n^{(k)} f(\mathbf{z}_n^{(k)}) \Big| \\
\leq&2\mathrm{E}^{(S)} \mathrm{E}_\sigma \sum_{k=1}^{K} w_k \frac{1}{N_k} \sup_{f\in\mathcal{F}} \Big| \sum_{n=1}^{N_k} \sigma_n^{(k)} f(\mathbf{z}_n^{(k)}) \Big| \\
=&2 \sum_{k=1}^{K} w_k \mathcal{R}^{(k)}(\mathcal{F}).
\end{aligned} \tag{108}
$$

By combining (105), (107) and (108), we obtain with probability at least $1-\epsilon$

$$
\sup_{f\in\mathcal{F}} \big| \mathrm{E}_{\mathbf{w}}^{(S)} f - \mathrm{E}^{(T)} f \big| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + 2 \sum_{k=1}^{K} w_k \mathcal{R}^{(k)}(\mathcal{F}) + \sqrt{\sum_{k=1}^{K} \frac{(b-a)^2 w_k^2 \ln(1/\epsilon)}{2N_k}}.
$$

This completes the proof. ■

C.3 Proof of Theorem 6.2

In order to prove Theorem 6.2, we also need the following result (see Bousquet *et al.*, 2004, Theorem 5):

Theorem C.3 *Let $\mathcal{F} \subseteq [a,b]^{\mathcal{Z}}$. For any $\epsilon > 0$, with probability at least $1-\epsilon$, there holds that for any $f \in \mathcal{F}$,*

$$
\begin{aligned}
\mathrm{E}f \leq& \mathrm{E}_N f + 2\mathcal{R}(\mathcal{F}) + \sqrt{\frac{(b-a)\ln(1/\epsilon)}{2N}} \\
\leq& \mathrm{E}_N f + 2\mathcal{R}_N(\mathcal{F}) + 3\sqrt{\frac{(b-a)\ln(2/\epsilon)}{2N}}.
\end{aligned} \tag{109}
$$

Again, we prove Theorem 6.2 by using Theorems C.2 and C.3.

Proof of Theorem 6.2 We only consider the result of Theorem C.2 in a special setting of $K = 2$, $N_1 = N_T$, $N_2 = N_S$, $w_1 = \tau$ and $w_2 = 1 - \tau$. Given a choice of $\tau \in [0, 1)$, denote

$$H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big) := \sup_{f\in\mathcal{F}} \big|\mathrm{E}_\tau f - \mathrm{E}^{(T)} f\big|. \tag{110}$$

By (9), we have

$$H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big) = \sup_{f\in\mathcal{F}} \Big|\tau \mathrm{E}_{N_T}^{(T)} f + (1-\tau)\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(T)} f\Big|, \tag{111}$$

According to Theorem C.2, we have for any $\xi > 0$,

$$\Pr\Big\{H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big) - \mathrm{E}^{(S)}\big\{H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big)\big\} \geq \xi\Big\} \leq \exp\left\{\frac{-2\xi^2}{(b-a)^2\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)}\right\},$$

which can be equivalently rewritten as with probability at least $1 - (\epsilon/2)$,

$$\begin{aligned}
&H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big) \\
\leq&\mathrm{E}^{(S)}\big\{H\big(\mathbf{Z}_1^{N_S}, \overline{\mathbf{Z}}_1^{N_T}\big)\big\} + \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)} \\
=&\mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\tau\mathrm{E}_{N_T}^{(T)} f + (1-\tau)\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(T)} f\Big|\right\} + \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)} \\
=&\mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\tau\big(\mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f\big) + (1-\tau)\big(\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(T)} f\big)\Big|\right\} \\
&+ \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)} \\
\leq&\tau\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f\Big| + (1-\tau)\mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(T)} f\Big|\right\} \\
&+ \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)} \\
=&\tau\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f\Big| + (1-\tau)\mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(S)} f + \mathrm{E}^{(S)} f - \mathrm{E}^{(T)} f\Big|\right\} \\
&+ \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)} \\
\leq&\tau\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f\Big| + (1-\tau)\mathrm{E}^{(S)}\left\{\sup_{f\in\mathcal{F}}\Big|\mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(S)} f\Big|\right\} \\
&+ (1-\tau)\sup_{f\in\mathcal{F}}\Big|\mathrm{E}^{(S)} f - \mathrm{E}^{(T)} f\Big| + \sqrt{\frac{(b-a)^2\ln(2/\epsilon)}{2}\left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S}\right)}
\end{aligned}$$

$$
\begin{aligned}
=&\tau \sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f \right| + (1-\tau)\mathrm{E}^{(S)} \left\{ \sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(S)} f \right| \right\} + (1-\tau) D_{\mathcal{F}}(S,T) \\
&+ \sqrt{\frac{(b-a)^2 \ln(2/\epsilon)}{2} \left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S} \right)}. \qquad (112)
\end{aligned}
$$

According to Theorem C.3, for any $\epsilon > 0$, we have with at least $1 - (\epsilon/2)$,

$$
\sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_T}^{(T)} f - \mathrm{E}^{(T)} f \right| \leq 2\mathcal{R}_{N_T}^{(T)}(\mathcal{F}) + 3\sqrt{\frac{(b-a)\ln(4/\epsilon)}{2N_T}}. \qquad (113)
$$

Next, according to (23), we have

$$
\begin{aligned}
&\mathrm{E}^{(S)} \left\{ \sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_S}^{(S)} f - \mathrm{E}^{(S)} f \right| \right\} \\
=&\mathrm{E}^{(S)} \sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_S}^{(S)} f - \mathrm{E}'^{(S)} \{ \mathrm{E}'^{(S)}_{N_S} f \} \right| \\
\leq&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \sup_{f\in\mathcal{F}} \left| \mathrm{E}_{N_k}^{(S)} f - \mathrm{E}'^{(S)}_{N_k} f \right| \\
=&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \sup_{f\in\mathcal{F}} \left| \frac{1}{N_S} \sum_{n=1}^{N_S} \left(f(\mathbf{z}_n^{(S)}) - f(\mathbf{z}'^{(S)}_n) \right) \right| \\
\leq&\mathrm{E}^{(S)} \mathrm{E}'^{(S)} \mathrm{E}_\sigma \sup_{f\in\mathcal{F}} \left| \frac{1}{N_S} \sum_{n=1}^{N_S} \sigma_n \left(f(\mathbf{z}_n^{(S)}) - f(\mathbf{z}'^{(S)}_n) \right) \right| \\
\leq&2\mathrm{E}^{(S)} \mathrm{E}_\sigma \sup_{f\in\mathcal{F}} \left| \frac{1}{N_S} \sum_{n=1}^{N_S} \sigma_n f(\mathbf{z}_n^{(S)}) \right| \\
=&2\mathcal{R}^{(S)}(\mathcal{F}). \qquad (114)
\end{aligned}
$$

By combining (110), (112), (113) and (114), we obtain with probability at least $1 - \epsilon$,

$$
\begin{aligned}
\sup_{f\in\mathcal{F}} \left| \mathrm{E}_\tau f - \mathrm{E}^{(T)} f \right| \leq&(1-\tau) D_{\mathcal{F}}(S,T) + 2(1-\tau)\mathcal{R}^{(S)}(\mathcal{F}) \\
&+ 2\tau\mathcal{R}_{N_T}^{(T)}(\mathcal{F}) + 3\tau\sqrt{\frac{(b-a)\ln(4/\epsilon)}{2N_T}} \\
&+ (1-\tau)\sqrt{\frac{(b-a)^2 \ln(2/\epsilon)}{2} \left(\frac{\tau^2}{N_T} + \frac{(1-\tau)^2}{N_S} \right)}.
\end{aligned}
$$

This completes the proof. ■